Kinetic models of metabolism: model construction, model analysis and biotechnological applications

Adrien Henry

UNIVERSITÉ SORBONNE PARIS CITÉ
UNIVERSITÉ PARIS DIDEROT

École Doctorale: Frontières du Vivant (ED 474)

Laboratoire: Génétique Quantitative et Évolution - Le Moulon

Doctorat

Spécialité: Biologie des systèmes

Présenté par:

Adrien Henry

Sujet:

Modélisation cinétique du métabolisme: construction du modèle, analyse et applications biotechnologiques

Kinetic models of metabolism: model construction, model analysis and biotechnological applications

Soutenue le 17 décembre 2015

Jury:

Daniel Kahn	Rapporteur
Jean-Pierre Mazat	Rapporteur
Olivier Martin	Directeur de thèse
Hidde De Jong	Examinateur
Philippe Nghe	Examinateur

Remerciements

Si cette thèse porte mon nom en tant qu'auteur, il serait illusoire de croire qu'elle eut pu être écrite par mon seul travail et dans un isolement le plus total. Je tiens donc à accorder le crédit qui leur est dû aux personnes présentes à mes cotés durant ces trois dernières années, que ce soit pour me conseiller, m'orienter ou bien simplement être là pour partager des les bons moments.

Je voudrais commencer par remercier Olivier, mon superviseur, pour m'avoir fait confiance dès le début de ma thèse de master. Tout au long de mon doctorat, il m'a apporté un nombre incalculable d'idées nouvelles lorsque j'étais bloqué, et il a su me guider dans ce projet jusqu'au bout. Je suis d'autant plus reconnaissant vis à vis de sa direction que ses orientations m'ont toujours été suggérées et jamais imposées. La liberté et la confiance qu'il m'a accordé durant les derniers mois de thèse ont été très précieuses, surtout qu'il n'était pas évident que je réussisse à tout finir à temps à ce moment là. Grâce à Olivier, j'aurais vraiment appris ce qu'est la recherche , ne pas se contenter d'un résultat mais toujours le questionner, l'approfondir. Ce besoin d'aller jusqu'au bout des choses est ertainement ce que je garderai de plus précieux après l'obtention de ma thèse.

En commençant ce projet, j'étais novice en biologie et plus particulièrement dans le domaine du métabolisme. Je dois beaucoup à toutes les personnes que j'ai rencontré durant les différentes conférences, et qui m'ont conseillé et critiqué. Je pense notamment à Ron Milo, qui m'a accueilli dans son groupe au Wiezmann Institute, à Avi Flamholtz, son étudiant ainsi qu'à Wolfram Lierbermeister de l'Université Charité de Berlin. Leurs modèles ont été pour moi une grande source d'inspiration. Je n'oublie pas non plus Grégory Batt et Armel Guyonvarch, avec qui nos discussions lors de mes comités de thèse ont enrichi ma culture biotechnologique.

Sur la liste des éléments déterminants à la réussite de mon doctorat, je ne peux pas oublier le projet RESET(ANR-11-BINF-0005) qui a financé mes trois années de thèse. Au delà de l'aspect purement financier, les réunions du projet ont été une source de motivation. Se rendre compte de l'application concrète des modèles à de vrais organismes vivants a été très excitant. Beaucoup des idées que j'ai pu avoir lors de ces trois ans sont dues aux nombreuses échanges que j'ai pu avoir avec Delphine Ropers et Hidde de Jong.

Je tiens à dire un grand merci aux membres de mon jury. Tout d'abord à mes rapporteurs Daniel Kahn et Jean Pierre-Mazat, leurs critiques, toujours constructives, ont été très enrichissantes. Daniel m'a permis de développer mon travail sur les temps de relaxation. Jean-Pierre, que j'ai eu la chance de rencontrer durant une conférence « Modelling Complex Biological Systems in the Context of Genomics », m'a beaucoup appris et son rôle de consultant sur le métabolisme à été précieux. Je remercie aussi Hidde de Jong et Philippe Nghe d'avoir accepté d'examiner ma thèse.

J'exprime aussi ma gratitude envers les membres du laboratoire de génétique quantitative et évolutive du Moulon qui m'ont accueilli au sein de leur unité, en particulier les membres de l'équipe BASE. Les discussions et interactions que j'ai pu avoir avec Aurélie Bourgeais, Christine Dillman, Judith Legrand, Adrienne Ressayre, et Dominique de Vienne ont été très instructives. J'ai beaucoup apprécié leurs commentaires et critiques lors des présentations que j'ai faites ou mes répétitions de soutenance. Merci également à Rozenn le Guyader et Valérie Lespidas pour toute l'aide qu'elles m'ont fourni.

D'un point de vu moins académique mais au moins aussi important je pense que le soutien de mes amis lors de ces trois dernières années a joué un rôle crucial dans l'aboutissement de ma thèse. Pourvoir passer du bons moments avec eux m'a permi de décompresser, de me détendre et de relativiser quand les travaux de thèse n'avançaient pas aussi bien que prévu. Bien qu'il n'en n'est fait mention nulle part dans ce manuscrit, durant cette thèse j'aurais passé beaucoup de temps à voyager, écouter de la musique, aller au cinéma, faire du sport (pas évident ces derniers mois), aller prendre un verre, etc. Ces activités auraient été beaucoup moins intéressantes sans la comagnie des personnes avec qui je les ai faites. Cela inclus beaucoup de monde, et je voudrais dire merci à chacun, avec notamment une reconnaissance toute particulière pour mes colocataires Jean-Christophe et Matthieu. De même que J'ai beaucoup apprécié le temps passé avec les amis du laboratoire, merci à Bubar, Christophe, Dorian, Cyril, Héloïse, Julie, Margaux-Alison, Sandra et Yasmine. Je voudrais encore remercier trois personnes avec qui j'ai souvent eu l'occasion d'aller manger un morceau ou boire le thé, Anncharlott, Livia et Samuel.

Je voudrais dire aussi un grand merci à tous mes amis de longue date que je ne vois plus très souvent mais que les revoir est toujours un vrai plaisir, je pense notamment à Alexis, Antoine, Pierre-Luc, Lucie, Micha et Rachel. Merci aux amis que je garde depuis le magistère, Benjamin, Laetitia, Pierre, Matthieu, Maud et Yannick. Je remercie Antoine en repensant avec beaucoup de plaisir à tous les bons moments au lycée, ce fut un plaisir de te retrouver en région parisienne et de continuer de nouveau à partager des moments ensemble durant ces dernières années.

Pour terminer, je voudrais consacrer ce dernier paragraphe pour exprimer toute ma reconnaissance aux personnes qui me sont particulièrement chères, et un merci ne sera probablement pas suffisant pour exprimer toute ma gratitude. Pour commencer, je voudrais dire un énorme merci à Lucie, ton soutien a été plus que précieux. Merci pour ta patience et ta très grande compréhension. Tu as toujours respecté ce projet personnel et tu m'a aidé à aller jusqu'au bout dans les moments de doutes. Ensuite je voudrais remercier mes parents d'avoir respecté mes choix et de m'avoir permis de suivre la voix que je voulais, merci aussi à eux de m'avoir soutenu durant toutes ces années. Je remercie également et fortement mes frères, pour tous les bons moments qu'on a passé ensemble à chaque fois qu'on se revoyait en Bretagne. Pour terminer je voudrais dire un grand merci à chacun de mes grand-parents, ça a très souvent été une source de motivation de savoir que vous étiez fiers de ce que j'entreprenais même si j'ai dû m'éloigner. Remercier ma copine et ma famille est pour moi le plus important de cette longue liste, la chaleur de votre présence et votre soutien ont toujours été là pour assurer la sécurité et le confort nécessaire afin de pouvoir entreprendre tous mes projets.

Résumé

Cette thèse décrit une méthode pour développer un modèle du métabolisme carboné central chez la bactérie *Escherichia coli* afin de tester une stratégie de bio-ingénierie sur une souche pour laquelle la machinerie d'expression génique(GEM) est contrôlable. L'idée est de réorienter la machine cellulaire depuis sa croissance vers la production d'un composé industriellement intéressant. La bactérie ainsi contrôlée ne va plus maximiser sa croissance, ce qui rend le cadre de la "Flux balance analysis" inapproprié pour la modélisation; un modèle cinétique lui est préféré. Étant donné le nombre important de réactions présentes dans le réseau, un pipeline a été mis en place pour produire automatiquement les lois cinétiques à partir des stœchiométries de réaction. Dans ce contexte, une description précise des mécanismes de réaction est impossible ce qui m'a poussé à choisir des modélisations de type "convenience kinetic" pour les réactions réversibles ou "Michaelis-Menten" pour les irréversibles; dans les deux cas les réactants sont supposés indépendants. L'ajustement des paramètres cherche à s'accorder au mieux avec des valeurs à l'état stationnaire de flux et concentrations, des distributions *a priori* de paramètres construites à partir de la littérature ainsi que des données de dynamique pour des traceurs. La thèse met en avant l'importance d'intégrer ces dernières et décrit les différents temps qui caractérisent un tel système, notamment le temps de relaxation n'est pas toujours celui le plus lent. Pour finir, le modèle optimisé est utilisé pour montrer qu'inhiber le GEM permet d'augmenter le rendement pour la production d'un métabolite cible.

Mots-clés: Biologie des systèmes, Métabolisme, Modélisation, Dynamique, Bio-ingénierie, *Escherichia coli*

Abstract

This thesis shows how to build a kinetic model the central carbon metabolism of the *Escherichia coli* bacterium to test a bioengineering strategy where the gene expression machinery (GEM) is controllable. The idea is to reorient the machinery from growth to the production of industrially interesting compounds. Because this controlled bacterium will no longer maximize growth, flux balance frameworks are inadequate and instead a kinetic modelling approach is necessary. Given the large number of reactions included in the network, a pipeline has been built to automatically generate kinetic laws from reaction stoichiometries. In this context a precise description of the reactional mechanism is impossible and I use the convenience kinetic framework for reversible reaction or Michaelis-Menten for irreversible ones; both are derived assuming independent reactants. The parameter fitting searches for the model that matches best the steady state conditions of concentration and flux, prior distributions for parameters built from literature data, and time course data for tracers. The thesis highlights the importance of including these time courses and of understanding the different characteristic times in such systems, the standard relaxation time not always being the longest characteristic time. Lastly, the optimised model is used to show that the yield of a target metabolite is increased by down regulating the GEM.

Key words: Systems biology, Metabolism, Modelization, Dynamics, Bioengineering, *Escherichia coli*

Contents

1

Introduction

1.1 Metabolic capabilities of living cells

An astonishing property shared by all living systems is their ability to use inert matter (and sometimes even very simple molecules) to produce sophisticated biomolecules that interact together and lead to reactive, adaptive and even reproductive behavior, the essence of what one considers as being alive. Life is a highly complex and self organised phenomenon where interactions arise on many scales, but all are based on physical or chemical processes with intricate feedbacks and regulation. Furthermore, there are always many many molecular species involved: it seems that Life cannot be achieved without a high level of complexity in its components. Typically, each of the many parts has its role in the "proper" functioning of the whole system; for instance, in bacteria peptoglycan molecules form arrays in the cell wall, enzymes catalyze reactions forming pathways, etc. Not only does a living system maintain itself out of equilibrium (equilibrium means death), but its self-organization may allow it to move, grow, send signals, self-replicate... Today, all biological systems are built from the fundamental module consisting of a single cell. In some living systems, the organism is limited to just one cell, and in others there may be a large number of cells organized so that it is the whole system which replicates. One is far from understanding how all the parts of such organisms interact and allow for reproduction, though prokaryotes (undergoing binary fission) involve certainly less complex processes than higher eukaryotes which undergo sexual reproduction.

Organization and out of equilibrium states in living organisms require a flux of incoming energy, as is also true for producing the biomolecules that compose individual cells. Both energy and building blocks of biomolecules are generated from external nutrient molecules thanks to (bio)chemical reactions, the collection of which is generally called the organism's metabolism. The function of metabolism is first to break down molecules through catabolism to generate small building blocks that can serve for the formation of larger molecules of interest to the cell. Catabolic chemical reactions may also degrade the nutrients all the way down to waste products which are excreted (examples include carbon dioxide and water); that "complete" catabolism is of interest if it releases energy; some of that energy is released as heat but the cell generally captures a fair amount of it in chemical form, often storing it in high energy bonds as in ATP or NADH. These metabolites can be used to drive reactions that are otherwise thermodynamically unfavorable like polymerization or to provide the energy for physical processes in the cell. Such processes include transport of vesicles via motors or the pumping of molecules across a membrane to counterbalance osmotic changes. Figure Fig. 1.1 shows a number of different scales in the organization and complexity that arise in unicellular organisms. On the lower scales, simple molecules serve as bricks for larger (bio)molecules; this is made possible through metabolism where simple compounds are transformed into precursors, precursors are used to make bio-blocks, bio-blocks are assembled into biopolymers, etc... Naturally there are many higher levels but these will not be studied in this thesis.

Across all living organisms, much of metabolism consists in the transformation of organic compounds (molecules made out of carbon) into other organic compounds. Why is carbon omnipresent in the molecules used in living systems? This "enrichment" is not due to the abundance of carbon in the environment provided by our planet since carbon is relatively rare. Indeed, carbon accounts for only 0.19% of the earth's crust. Looking to the atmosphere, CO_2 accounts there for only a 0.04% fraction (this fraction is increasing steadily as everyone knows). The explanation for why life is so anchored in carbon is chemical: carbon has a high propensity to make strong covalent bonds by electron pair sharing with itself or with other elements. Quantum mechanics teaches us that covalent bounds are more stable between lighter atoms since the quantum energy levels occupied by those bonds are lower and require more energy to be broken. Hydrogen, oxygen, and nitrogen are three other light elements that share with the carbon the ability to form covalent bonds to stabilize their electronic environment and are very well represented in biomolecules too. Carbon, hydrogen, oxygen and nitrogen can form respectively four, one, two, and three covalent bonds with other molecules by sharing their valence electrons; that allows them to form an important number of different molecules that are amply exploited in living systems.

Both inorganic and organic molecules are present naturally on earth, and inorganic like organic compounds can react together. If a reaction is exothermic, it can happen spontaneously. If instead it is endothermic, the reaction can still arise if different sources of free energy can be tapped [25] like light

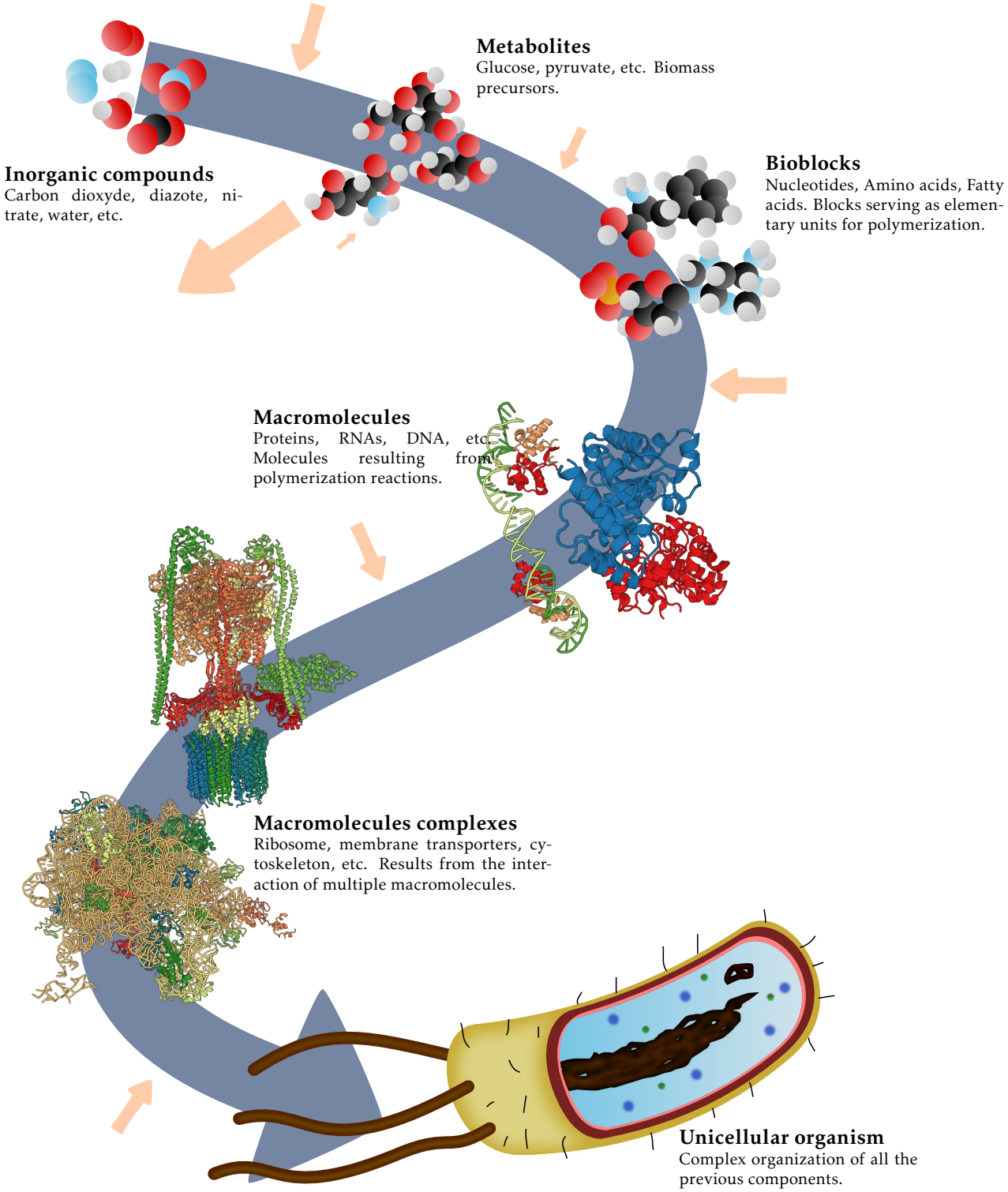


Figure 1.1: Different scales of complexity in a prokaryotic cell. To go from one level to the next generally requires the addition of energy, symbolized here by orange inward arrows. The source of energy for the cell originates mostly from the catabolysis of energetic metabolites such as sugars, cf. the outward orange arrows. Drawing from www.rcsb.org.

from the sun, temperature gradients arising in deep sea vents, etc. Living organisms can also produce non-organic molecules of course. Inorganic compounds are both produced and decomposed, and again when necessary, outside sources of energy can be exploited, as happens for instance in photosynthesis where synthesis of various metabolites is driven by sunlight.

1.2 Reactions and enzymes

A cell is capable of sustaining and renewing all of its components and to precisely organise internal processes in a spatio-temporal program. It is remarkable that higher level objects such as proteins and DNA are built according to very similar procedures across all organisms. For example, proteins are almost always made of the same 20 amino acids (AA) that are common to nearly all species, only a few cases of other AAs arise naturally. These AA are linearly assembled by polymerization using the very sophisticated machinery of ribosomes, machinery that varies very little from organism to organism. Similarly, DNA (respectively RNA) is produced from the same four nucleotides (in fact three of these are common across DNA and RNA, the fourth is thymine for DNA versus uracil for RNA) and again the polymerization of these biomolecules proceeds through polymerases that have much in common across the different domains of life. The building of all these macromolecules is highly deterministic, and the chemical reactions producing the bio-bricks or other metabolites are under strict control in all cells. Perhaps the very first forms of life had less organized and determistic processes, involving reactions with less specificity and even perhaps involving a significant amount of randomness. It has been argued that something similar to that kind of randomness still occurs today at the level of transcription: with the ENCODE project, it has been suggested that a large fraction of DNA is transcribed even though there are no obvious functions for the associated RNAs; such a picture seemed appealing to many, as rare molecular species provide a source for selection to work on. Even if that picture is ultimately maintained (many argue it should be abandoned), no such point of view seems to be mainstream today in the context of metabolism. Indeed, metabolic performance has been under strong selection for billions of years so it seems unlikely that chemical reactions in today's species would still lead to "random" metabolites. To add credence to this point, it seems likely that random reactions would tend to remove essential metabolites rather than produce new interesting ones.

Most biochemical reactions are naturally very slow if no catalyst is present. The slowness of metabolic reactions in the absence of catalysts is in fact a key factor that allows those reactions to be controlled, and thus for life to sustain itself. To bring this point home, recall that proteins naturally break down into peptides through spontaneous hydrolysis; if such reactions were much faster, it would never be possible to maintain a cell's integrity. Inversely, if synthesis reactions could not be speeded-up, it would not be possible to fight the natural decay processes. Fortunately, most reactions can have their rates go up by orders of magnitude in presence of enzymes (by up to factors of 10^{10}). Catalysis of biochemical reactions is almost always provided by dedicated proteins or proteic complexes in the cell, and these proteins are then called enzymes. Proteins are polymers of AAs and depending on their sequence they fold to produce a three dimensional structure. In the case of enzymes, this three-dimensional structure incorporates a region referred to as the active site, this site binding substrates and thereby lowering the activation energy of the reaction. Because of Arrhenius' law, reaction rates decrease exponentially with activation energies; it is thus possible to have rates of reactions go *up* by large factors if the enzyme's structure is right so as to lower activation energies. For most reactions of central metabolism, enzymes have been subject to natural selection for billions of years and so are now near "optimal". For any protein, its sequence of AAs is assembled by polymerization within the cell from a template encoded in RNA, itself obtained from a DNA template subject to mutation and thus evolution; the enzymes have been selected for their efficiency but also based on their production cost in term of energy and nutrients; such production costs are roughly proportional to the linear length of the enzyme. The important point here is that enzymes, since they have been selected for their efficiency, are typically highly specific to one reaction. Through a complex regulatory program, a cell can choose which enzymes to produce, hence which reactions or even pathways to turn on. An enzyme can catalyze only a certain number of reactions per second; as a consequence, reaction fluxes are typically limited by the number of (active) enzyme molecules in a cell. Regulating the rate of a reaction can be done by regulating the number of

molecules of its enzyme, but there are also other regulatory mechanism that we will mention in the next chapter. Overall, cells have generated a multitude of methods for controlling their metabolism, in far more subtle ways than using on-off switches.

In physiological cellular conditions, only the reactions that are catalysed by an enzyme can proceed at rates which matter to the cell. The putative random reactions that happen naturally run at a slow rate and thus do not compete much against enzymatically driven reactions. The importance of enzymes for biochemical reactions is so major that people indifferently refer to the reaction or to the enzyme with the same name. Properties and modeling of enzymatic reactions are described in more detail in chapter 2.

1.3 The central dogma of biology

We mentioned that each protein in a cell consists of a chain amino acids whose order is encoded in regions of the DNA, loosely referred to as genes. The monk Gregor Mendel realized in 1860 by crossing peas of different colors and shapes that information is transmitted stochastically from the parents to descendance, but that this transmission obeyed statistical laws that were *simple*. In 1953, James Watson and Francis Crick discovered the double helix structure of DNA [79] using diffraction data produced by Rosalind Franklin. DNA is a nucleic acid formed of two helices, each forming a backbone with attached nucleotides. These nucleotides allow the two helices to bind non covalently. DNA uses four nucleotides, adenine (A), cytosine (C), guanine (G), and thymine (T) and the sequences formed typically encode information that is exploited by the organism and which it transmits to its descendants. Part of the magic of DNA is its double helical structure allowing the unzipping of the two parts followed by faithful copying of each strand. That elegant feature explains why the discovery of Watson and Crick had such an impact (beauty in science is often the key to great discoveries). The nucleotides A,C,G,T are associated in pairs to bring together the two strands of DNA, but what is essential is that the bindings are reciprocal and specific: between A and T on the one hand and between C and G on the other.

In 1970, Crick proposed a framework [17] to understand how the information contained in DNA might be converted, so that a nucleotide sequence might uniquely determine an amino acid sequence and thus a protein as shown in Fig. 1.2. Interestingly, that conversion is not direct in its biochemical implementation. First a working copy of the information is produced by transcription, leading to a first product embodied in RNA and called the messenger RNA (mRNA). A large complex, the RNA-polymerase (RNA*pol*) is the (ATP-dependent) motor driving this transcription; it incorporates into the mRNA a G if the DNA has a C and reciprocally. Furthermore, RNA*pol* incorporates a uracil (U) nucleotide if the DNA has an A instead of the T one might naively have expected (indeed, if the DNA has a T, RNA*pol* incorporates an A). The mRNA sequence of nucleotides is then used as a template for producing the protein. This is executed by the ribosome complex, a huge machinery also powered by ATP. The mRNA is "read" three nucleotides at a time, so one such triplet is referred to as a codon. Each of the 64 codons can be thought of a unit of information, and corresponds to an instruction to start or stop the machinery or to add one (codon-dependent) specific AA to the polymerizing polypeptide to form the target protein. The start and stop sequence tells the ribosome where the coding part of the gene, i.e., the sequence coding for one protein, starts and stops. The mapping from codon to AA is referred to as the genetic code, and it is nearly universal, almost all organisms using the same code.

The central dogma of molecular biology, as described in the previous paragraph, is a simplified but relevant representation of the core processes of the cellular machinery. Since its original formulation where the information flowed unidirectionally from DNA to proteins, the dogma has been refined. For instance, one now knows that there are many RNAs which are produced but do not lead to translation (non coding RNAs). These regions are then referred to as non coding regions; their functions are extremely diverse, allowing for instance the cell to modify its genetic program or to protect against pathogens. There are also other regions of the genome which do not get transcribed but can play a role in the transcriptional machinery. For example, the DNA upstream of the transcription start side can affect rates of transcription by binding various specialized actors (transcription factors) which will affect the probability that an RNA*pol* will be recruited. Such DNA sites, called promoters, are of critical importance for the regulation of the cell since the affect gene expression (via transcriptional regulation).

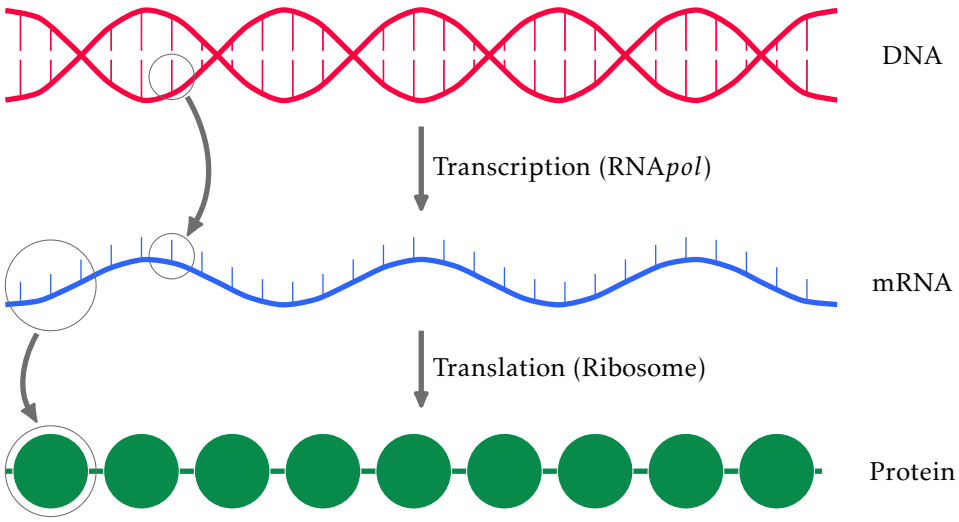


Figure 1.2: Central dogma of molecular biology. The DNA sequence (consisting of a string of A,C,G and Ts) of a gene is transcribed to form mRNA composed of of nucleotides A,C,G, and U (U playing the role T plays in DNA). This copy then serves for the translation process, taking the information in the codons encoded in the mRNA into a sequence of AA forming the primary structure of a protein. This translation arises thanks to the action of the ribosomal machinery that adds an amino acid to a polymerizing polypeptide for each codon (consisting of three successive nucleotides).

These regions are often targeted in bio-engineering because they can provide a way to repress or activate specific genes. Gene expression is often modulated at the post-translational level: a protein may undergo changes which affect its ability to execute its function. Such processes are extremely common in signaling cascades via phosphorylation of certain residues of the protein actors. These are all minor changes to the original central dogma. More profound changes have been added in the last 20 years because of the discovery of certain actions of proteins or RNAs on DNA itself; examples include jumping genes and epigenetic processes.

1.4 Metabolic networks: reconstruction

Since almost every reaction in an organism's metabolism is catalyzed by an enzyme, the first thing to do in order to study its metabolism is to search for all the enzymes present in the cell to obtain a complete list of all possible reactions that can arise. If indeed one knows that an enzyme is able to catalyze a reaction, its presence can be considered as indicative that the reaction is used. It is best for the complete stoichiometry of the reaction to be known exactly (substrates and products and associated proportions). In general one relies on knowledge in various organisms and extrapolates to new organisms using homology between protein sequences. In the last 20 years much of this kind of metabolic reconstruction work has been performed, to provide as exact as possible reaction lists in model organisms such as *E. coli* but also to extend these inferences to other less well studied organisms. These tasks are difficult and have required a lot of work, often involving large teams, but also have led to impressive successes [22].

The first metabolic network reconstructions were performed during the period 1985-1995, and focused on *Clostridium y* [61], *Bacillus subtilis* [62], and *Escherichia coli* [76]. At that time metabolic reconstruction relied on an extensive literature exploration to find evidence of the association between an enzyme and a reaction. But with growing numbers of metabolisms described, databases like MetaCyc [12], KEGG [44], or Brenda [68] have been developed. They contained information about gene sequences and the putative enzymatic functions of the associated proteins in a number of different organisms.

The more recent metabolic reconstructions are often based on a fair amount of automatic processes. High throughput methods provide extensive information about genomes, from which gene models can

be inferred. This allows one then to implement comparative genomics to search for orthologies. To do so, gene sequences are compared to already published anotated genomes; if the homology is high enough, it becomes realistic to transfer the annotated function from one gene to the other. This method is thus used to identify many putative enzymes in the system studied, but such inferred metabolic properties must be used with caution. A manual curation of the model may be still necessary to provide confidence in the extrapolation and to test whether the metabolic model is in agreement with the physiology of the organism studied. Such tests can be labourious because there are often orphan reactions or inversely enzymes whose function remains unclear. Unless there is a big stake, it is not possible to check that each reaction generated automatically is indeed realized in the organism. A simpler though less comprehensive approach consists in comparing the growth behavior of the organism on different media with what is predicted by the reconstructed metabolic network. Proteomic data can also help to know in which condition such and such an enzyme is present.

1.5 Context of the RESET project

This thesis stands on its own but it is nevertheless appropriate to consider the context in which it was designed. My three years of work were funded by the "Projet d'Investissement d'Avenir" (PIA) project entitled "RESET". That project is an effort involving experimentalists and theoreticians to modify *E. coli* cells for metabolic engineering but using a non metabolic strategy, an approach which thus is quite unusual and innovative. Within standard metabolic engineering approaches, one searches for the best set of enzymes to over-express or to knockout (to enhance or suppress reaction fluxes in pathways of interest). Instead of that, RESET adopts an indirect approach aimed at affecting the overall gene expression machinery (cf. the section on the central dogma of molecular biology). It consists in globally controlling gene expression rather than focusing on a few enzymes. The motivation behind adopting such a global approach is that by preventing the cell from using its building blocks (AA, nucleotides, etc.) for growth, metabolic fluxes should be reoriented, possibly to pathways of interest.

Partners in the project have developed an *E. coli* strain in which the gene operon coding for the β and β' subunits (rpoBC operon) of the RNA*pol* is under control of the metabolite IPTG. Specifically, the promoter of rpoBS is replaced by the promoter of the operon lactose. The lac operon [38] is composed of three genes lacZ, lacY and lacY and is inhibited by the protein coded by a fourth gene, lacI. The protein encoded by lacI has a high affinity for the lactose operon and is constitutively expressed. When lactose enters the system, it is converted to allolactose which binds to LacI protein and thus prevents the inhibition of the lac operon. The molecule IPTG has the same property of binding to the LacI protein as allolactose but it also has the advantage of being more stable which allows a fine-tuning of its concentration in the medium. In the absence of IPTG, the rpoBC is inhibited by lacI. As a result, since the core subunits β and β' are no longer transcribed, no new RNA*pol* can be produced which means that transcription rates inevitably decay, and in fact for all genes of the genome.

In the absence of IPTG, there is less renewal of the mRNA pool; note that in bacteria, mRNAs are quite unstable, they get degraded fast and have an average lifetime of 5 min [10]. Thus, if one washes a cell suspension, the IPTG will be diluted away, and so the presence of mRNAs in the cell will depend on the (decreasing) numbers of RNA*pol*. With regard to proteins, their half-life of protein is much longer, typically on the order of an hour or more, which means that even after transcription has been stopped, the protein quantity will remains more or less constant except for dilution effects, if the cells continue to grow. It has been shown experimentally that the transformed cells keep growing after removal of IPTG, but at a slower rate. The central idea of the RESET project is to stop the transcription. A direct consequence will be that the "consumption" of nucleotides and also AA will go down. Indeed, the mRNAs will be produced at a decreasing rate, so the pool of nucleotides should rise, and the lower rate of translation should also lead to an increase in the pool of AA. The accumulation of nucleotides and amino acids should inhibit their biosynthesis pathways while leaving the uptake of carbon and central metabolism available for other purposes. In RESET, a proof of concept of this idea is being tested via the production of glycerol. The figure Fig: 1.3 presents a schematic picture of this strategy. The name of RESET originates from the restart of the cell's gene expression machinery by addition of IPTG to the medium. This resetting is necessary since once the degradation of house keeping proteins becomes too

severe, one has to rescue the cells by letting them get back to their maintenance activities which require a functioning gene expression machinery.

Figure 1.3: The strategy of the RESET project. A: IPTG is present in the medium so the cell functions normally, a large fraction of the the flux through central metabolism is dedicated to the production of amino acids and other building blocks essential for growth. B: The cells are washed to remove IPTG; then the gene expression machinery is inhibited; quickly, the requirements for amino acids and nucleotides are lowered, leading to a reallocation of the metabolic fluxes for other purposes such as the production of other compounds of interest.

As mentioned earlier, the RESET project involves both experimentalists and modelers. The modeling part's objective is to couple (i) a mathematical model for the gene expression machinery, developed at INRIA Grenoble by Delphine Ropers, and (ii) the kinetic model of central carbon metabolism developed in this thesis. The modeling approaches should help us understand the reallocation of fluxes when RNA*pol* synthesis is turned on and off, and may help orient certain choices for optimizing the *E. coli* strains.

1.6 Flux balance analysis: a powerful modeling framework at steady state

There exists several types of *in silico* models that give insights about the distribution of the fluxes of molecules through each reaction of a metabolic network. Before explaining why those frameworks are not appropriate for being used in the RESET project, I need to explain them a bit. The most common framework is based on flux balance analysis (FBA). That approach has been very useful for predicting metabolic capabilities of different organisms, thanks in part to the rapid increase in knowledge of the genomes of new organisms and their associated annotations. The overall method has been turned into a very powerful tool B. Palsson and his group since 1992 [67]. There have been a great number of reviews that describe FBA [46, 64], but the basic concepts are quite simple and I now briefly explain them.

A chemical reaction is the transformation of one or more substrates to one or more products. The key point of FBA is that these reactions preserve mass (all atomic species). For example the reaction aldolase transketolase A, *i.e.*,

$$\begin{aligned} E4P + X5P &\rightleftharpoons F6P + GAP \\ C_4H_7O_7P + C_5H_9O_8P &\rightleftharpoons C_6H_{11}O_9P + C_3H_5O_6P \end{aligned}$$

transforms the substrate E4P and X5P (left side) to F6P and GAP (right side). It is easy to see that this reaction conserves all atoms: on both the left and right sides, there are 9 carbons (C), 16 hydrogens (H), 15 oxygens (O) and one phosphorus (P). The convention is to attribute a number, called the stoichiometric number, to every metabolite involved in a reaction. The stoichiometry accounts for the proportion of the metabolites that enter the reaction. Then for each atomic species, one has a conservation law that

can be written in terms of these stoichiometric coefficients and the number of occurrences of the atom in each metabolite. By convention, the stoichiometric coefficient is negative for substrates and positive for products. Mass conservation of the considered reaction then leads to "balance" equations for each atom:

$$\sum_i s_i n_i = 0$$

where n_i is the number of times the given atom occurs in metabolite i.

FBA is a constraint-based approach appropriate for describing the possible fluxes when the system is at a steady state. Let us begin by describing the dynamics of metabolite concentrations in a linear algebra framework. A cell metabolism is a network containing n metabolites connected by m reactions. The concentration vector $\vec{C} \in \mathbb{R}^n$ has dynamics which can be written in terms of the fluxes through the different reactions. Let $\vec{R} \in \mathbb{R}^m$ be the vector describing the rate of transformation per second for the m reactions. Then one has

$$\frac{d}{d\,t}\vec{C} = S\vec{V} \tag{1.1}$$

where $S \in \mathbb{R}^{n\times m}$ is the stoichiometry matrix. Thus the j^{est} column corresponds to the y coefficients of the metabolites involved in the reaction j (the coefficient vanishes if the metabolite is not involved in the reaction).

The reaction fluxes in $\vec{V}$ depend on kinetic laws, and so depend in particular on the concentrations of the metabolites in the network. These fluxes are thus in general time-dependent. The fundamental assumption of the flux balance analysis is to consider only networks at the steady state. Then the concentrations are also time-independent as there is as much flux consuming a metabolite as flux producing it. Steady-state conditions are easily ic in experiment; this occurs for instance in chemostats where environmental conditions are kept fixed. But it is also a good approximation in batch cultures because the population growth is slow compared to metabolic time scales and so one is in a quasi steady-state regime. Under such steady-state conditions, Eq. 1.1 becomes

$$S\vec{V} = 0 \tag{1.2}$$

and it is relatively easy to solve this linear set of equations. However the number of reactions in metabolic networks it typically lower than the number of metabolites. As a result, the system of equations is under-determined and instead of one unique solution for the fluxes $\vec{V_0}$, one has a high-dimensional space of solutions, in fact the dimension can go up to several hundred. Therefore, before solving such a system, one tries to add some physiological knowledge to constrain the fluxes with an upper(u_j) and lower(l_j) bound for every flux (v_j):

$$l_j \leq v_j \leq u_j \tag{1.3}$$

These types of constraints help to reduce the space of solutions but do not lead to a unique solution for the fluxes in the system. To choose which is the "best" set of fluxes in this space, one needs to have an *a ic* objective function that the optimal flux vector should verify. In practical applications of FBA, one generally consider that evolution has led to maximization of growth rate, so this is the objective function which is translated into maximizing the flux towards biomass production. Specifically, the rate of production of the bio-blocks and/or energy is maximized. In practice, one imposes specific proportions for all these bio-blocks through the composition of the organism. These proportions thus depend on the detailed composition of the cell in RNA, DNA, proteins, lipids, etc. The final step of FBA, to obtain the optimal flux, involves setting for instance the influx of nutrients. Indeed, without such an additional constraint, by linearity of the system, any rescaling of a solution is also a solution. Thus FBA provides information on relative fluxes and thus yields, but is not able to give insights into actual influx. As a result, it is not truly predictive of growth rates, only relative growth rates are accessible.

The FBA framework with the inclusion of an objective function has had a large impact in metabolic modeling. That impact of course has been made possible by the investment of many people in network reconstruction for all sorts of prokaryotes, and there have been thus many tests of the associated predictions. One of its key features is that it is essentially parameter free: all predictions are in principle

amenable to first principle computations. The comparison of *in silico* fluxes with measured fluxes has been used to confirm or infirm hypotheses on the presence or absence of certain enzymes for which homologies are uncertain. These methodologies have helped the recontruction of major genome scale models in more complex organisms such as *Saccharomyces cerevisiae* [20, 26] and even human [19]. The FBA framework is also extensively used for searching for optimal sets of mutations that might improve metabolic yield without impacting too much the system's fitness [11]. On the contrary, impacting a pathogen's fitness by targeting its metabolism is a useful strategy in drug discovery [39] and FBA is a powerful modeling tool to provide candidate targets.

Flux balance analysis also has certain disadvantages. A first example is the difficulty of FBA to account to quantitative changes of the proteome. FBA can theoretically model the addition or removal of reactions but is poorly adapted if the goal is to understand the consequences of increasing or decreasing enzyme concentrations. Formalisms like rFBA (for regulated FBA) [16] or other generalizations try to modulate the flux constraints of Eq. 1.3 to account for the changes in the environment or in enzyme concentrations but that comes at the cost of introducing kinetic parameters. To circumvent that, one may empirically add bounds to the rates v_j, in particular via the upper bounds u_j, but the associated predictions are not very reliable. Consequently extensions of FBA loose the great advantage of FBA of not requiring many parameters.

The RESET project aims at modeling variation in the metabolism when various (unnatural) manipulations are applied to the growth machinery of the cell. As a result, both the steady-state assumption and the optimization selection principle at the heart of FBA make it unadapted to the RESET objectives. There exist a method called dynamical FBA [52] that tries to account for slow changes in the environment by modeling the nutrient uptake with kinetic equations and updating the optimal fluxes regularly via a succession of steady states. This method inspired the technique I use to model the biosynthesis from metabolic precursors to bio-blocks in my model. However it remains inappropriate for the RESET framework because it still uses an objective function which can only be justified on evolutionary time scales. Maximizing growth rate is a phenomenon that occurs not after a time of adaptation of metabolism or gene expression but on time scales of many generations. There is thus no reason for the cells constructed in the RESET framework can be expected to follow a predefined optimization principle. These obstacles thus pushed be to turn toward a fully kinetic approach for *E. coli* 's central carbon metabolism (CCM), that part of metabolism which takes nutrients with carbon (such as glucose) and metabolize them to biosynthetic precursors.

1.7 Why not use an existing kinetic CCM model?

Kinetic models describe every reaction rate in a metabolic network via kinetic laws which depend on the concentrations of the metabolites reacting and of the molecules affecting the reactions. The concentrations evolve in time according to these reaction rates (determining the fluxes) because fluxes are sources and sinks of metabolites, cf. Eq. 1.1. The difference with FBA is that the reaction rates are explicit and depend on concentrations (FBA does not follow concentrations of metabolites, nor of enzymes). Thus kinetic modeling is far more challenging than building an FBA model: one requires the knowledge of the mechanisms that will determine the kinetic laws and also the parameters that characterize these laws. Generally speaking, obtaining these parameters is the major stumbling block preventing the development of successful kinetic models. At first, kinetic models were mainly used to describe specific parts of certain metabolic systems [13] and then extended to larger scale models including principally the central carbon metabolism [14, 43, 63]. However, these large kinetic models have had hardly any concrete applications other than a proof concept for fitting dedicated experimental data.

The parameter fitting of kinetic models is a huge challenge that has limited the development of such models, justifying why FBA approaches are so often preferred. Some parameters have been quantified and are available in the literature but those data are very sparse. An additional difficulty is that measurements of these parameters can be condition dependent. This difficulty tends to drive one to discard components of the models that are not mandatory, reducing network size and simplifying the reaction laws. A frequent example is the use of modeling where reactions are considered to be irreversible. This leads to models with fewer parameters but the approximation is relevant only for those reactions with

a very favorable thermodynamics, meaning that although products may be reconverted into substrate, it typically does not happen under physiological conditions. This approximation allows one to move forward for the purpose of building the model (a strategy often implemented in models tackled so far). More generally, such short cuts allow one to nevertheless propose a model which reproduces experimental measurements. Often these involve time series after introducing a pulse for instance of glucose, and will compare the behavior of a wild type to a strain having one or more genes knocked-out [43]. These models have enough parameters to allow for good agreement with the data set used for calibration but because of their ad-hoc choices, they generalize poorly and cannot adequately predict behavior in other experimental conditions without recalibration. For example they will not model both glycolysis and gluconeogenesis since some of the associated reactions are taken to be irreversible. Indeed, such a choice isolates modules along the pathway and prevent the upper modules from sensing an increase in concentration in the lower modules. In RESET, the environmental conditions change a lot after the arrest of the gene expression machinery so it is important that the metabolism be able to sense an increase in metabolite concentrations downstream. From a practical point of view, the model I have built needs to agree with steady states for reference data; it turns out that the sensing of the product by the reaction produces models with greater stability and leading more reliably to the steady state than models where the reactions can be irreversible.

A second difficulty in building kinetic models is the mechanistic description of the way in which an enzyme acts on its substrates and products. Furthermore, other factors may influence the rate laws, e.g., many biosynthetic pathways allow for regulation of un upstream enzyme by downstream products. The mathematical description of such laws leads to further parameters for which hardly anything is known experimentally. For pathways like the central carbon metabolism, they are qualitatively known and are generally included in published models. However, this fine level of description of interactions or regulatory control means more complex models with additional parameters. For example, in the case of allosteric regulation, an enzyme can be in different states depending on the concentration of an effector; the standard description [55] for such regulation involves new kinetic parameters for each of the states of the enzyme. With the ambition of building a general model of the central carbon metabolism that can respond to a large variety of conditions, such level of detail is inappropriate for the first stab at this challenging problem. In view of the lack of maturity of this field, the best I can hope for at the present time is the ability to provide a model giving qualitatively correct behavior for the transition between different conditions and in particular the ones relevant for the RESET project. Therefore, implementing refined descriptions of enzyme activity would contribute more to the complexity of the model (and its adjustment) than to the reliability of its predictions. The same argument applies for the description of enzyme saturation in substrate and product: a phenomenological approach with saturation but no real mechanistic encoding seems the most appropriate level to use in my model building.

Another factor that motivated me to build my own model is that often the kinetics represented in published models focus on the first moments after a perturbation is applied. This implies a precise description of the enzyme mechanism which I have already rejected but also it does not require any convergence to a steady state at long times. In particular, it is common practice in kinetic modeling papers to describe the time-dependence of metabolites, like cofactors that are involved in multiple reactions, using an ad-hoc function such as a polynomial of time [14] and see if the the other metabolites behave in agreement with the experimental measurements. Such frameworks are clearly not designed to include a steady state behavior. Specifically, the imposition by hand of a time-dependence for cofactors or other metabolites prevent convergence to a steady state; an unfortunate consequence is that the long time extrapolation of the kinetic model will often lead to vanishing or diverging concentrations. A last problem I discovered and which is presented in a later chapter is that some of these published kinetic models converge toward a steady state but with surprisingly high – and unrealistic – characteristic times [69], discrediting in effect the model's validity.

In conclusion, although there already exists a number of different kinetic models connected to the CCM, they are not suitable for the context of the RESET project. Furthermore, as justified by a number of arguments I presented above, it is not sensible to build a too detailed description of the different reactions and their regulations over different sets of condition. I therefore decided to use a partially coarse-grained approach for the description of the different reactions; this strategy maintains sufficient

parameter identifiability and hopefully does not sacrifice too prediction power. Ultimately, in future work, as data sets improve, some of these restrictions can be lifted. Perhaps the main take-home message is that I have provided a very systematic approach for building kinetic models of metabolism. The entry point is the metabolic network's topology while the end result, namely the calibrated kinetic model, depends on exploiting experimental measurements of systemic quantities.

1.8 Outline of the thesis

This thesis had as objective the construction of a kinetic model of *E. coli* 's central carbon metabolism. The scientific goal is to use such models to understand how a perturbation of the gene expression machinery can impact the production (flux) of a metabolite of interest, and specifically glycerol in the case of the RESET project that funded my work. Keeping this objective in mind, I propose a systematic and automatic approach to build kinetic models of any metabolic network of known stoichiometry. The thesis is organized as follows.

First I will describe how to generate qualitatively sensible reaction rate laws from the knowledge of the stoichiometry and of the y for reactions with any number of substrates and any number of reactions. Since kinetic y requires data to estimate unknown parameter values and test a model's relevance, I will also overview the experiments used in order to collect these kinds of data.

The next chapter arose from the observation that characteristic times in metabolic models can be surprisingly long. I will present the factors impacting characteristic times for a linear metabolic toy model based principally on theoretical methods, and then I will determine the values of these characteristic times in a number of kinetic models published and available in the database Biomodels.

The fourth chapter describes the actual model development. I will present in particular a framework I used to assess a model's goodness of fit. I will also explain the procedures I used to optimize the parameters set so as to fit as well as possible data from the literature and priors extracted from similar systems.

The final chapter presents the optimized model and the way I calculate the confidence intervals for each of its parameters. I explain the advantages of the different choices of measures used to optimize the parameters. This part also presents how my framework allows to compare predicted (unknown) parameter values to expectations as encoded in the prior distributions. I finish by presenting an *in silico* experiment that supports the RESET strategy for metabolic engineering while showing potential limitations.

Lastly, I close this thesis with a conclusion chapter in which I provide my outlook on this work.

2

Development of a kinetic model for *E. coli*

To build a kinetic model of a metabolic networks, one must specify the dependence of each reaction flux on the concentrations of the metabolites. Such rate laws reflect the frequency at which the substrates and enzymes meet and successfully produce the reaction leading to the products. In this section I will present the main rate laws used when modeling the dynamics of metabolic reactions. The properties of each approach will be listed in order to motivate choices for the methodology that fits best our need to have a sytematic method to generate a kinetic model even when the precise kinetics of the reactions are not known in detail.

2.1 Modeling kinetic reaction by convenience kinetic rate laws

2.1.1 Law of mass action

A chemical reaction is the conversion of a set of a molecules, the substrates, into another set of molecules, the products. The first quantitative description of such a conversion is historically associated to Guldberg and Waage (1864) [30] via the law of mass action. In that formalism, the rate of the reaction in the forward direction is directly proportional to each substrate concentrations. When the products of the reaction are present, they may also react together in an analogous manner to regenerate the substrates. (Most biochemical reactions of interest are reversible.) The global law for the rate accounts for this reverse flux by subtracting it from the forward flux. For illustration, let us consider a reaction having two substrates and two products: $S_1 + S_2 \rightleftharpoons P_1 + P_2$. The mass action law for the flux is then:

$$v = k^+ s_1 s_2 - k^- p_1 p_2 \tag{2.1}$$

Here v is the net flux (per unit volume) of the reaction, $k^+ s_1 s_2$ is the forward flux while $k^- p_1 p_2$ is the backward flux. The concentrations of S_1, S_2, P_1, P_2 are labelled respectively s_1, s_2, p_1, p_2 and are expressed in mol/L. k^+ and k^- are the affinity contants expressed in $mol^{-\alpha}s^{-1}$ and $mol^{-\beta}s^{-1}$ where α and β are the molecularity of the forward and backward reactions. In the present example $\alpha = \beta = 2$ since both the forward and backward reactions are bimolecular processes.

2.1.2 Thermodynamics

The thermodynamics of a reaction characterizes the relative importance of the backward flux as compared to the forward flux. Its knowledge tells us about the spontaneous direction of the total flux. For each metabolite involved in a reaction, one defines the Gibbs energy $\mu_i = \mu_i^0 + RT\log(c_i/c^0)$ where μ_i,R,T, and c^0 are respectively the standard Gibbs energy in the standard condition, the molar gaz constant, the absolute temperature in Kelvin), and the standard concentration of 1 $mol\,L^{-1}$ to work with dimensionless quantities. The reactional Gibbs energy change is defined as the change in energy for the system when one *mole* of substrate is converted. In the previous example, $S_1 + S_2 \rightleftharpoons P_1 + P_2$, the reactional Gibbs energy is

$$\begin{aligned}
\Delta G_r &= \mu_{P_1} + \mu_{P_2} - \mu_{S_1} - \mu_{S_2} + (\dots) = \overbrace{\mu_{P_1}^0 + \mu_{P_2}^0 - \mu_{S_1}^0 - \mu_{S_2}^0}^{\Delta_r G^0} + RT\log\overbrace{\left(\frac{p_1 \cdot p_2}{s_1 \cdot s_2}\right)}^{Q} \\
\Delta G_r &= \Delta G_r^0 + RT\log(Q)
\end{aligned} \tag{2.2}$$

From the sign of $\Delta_r G$ one gets the spontaneous direction of the flux:

- $\Delta_r G < 0$: Flux in the forward direction. $Q < e^{-G_r^0/RT}$
- $\Delta_r G = 0$: Reaction at equilibrium, no flux. $Q = e^{-G_r^0/RT}$
- $\Delta_r G > 0$: Flux in the backward direction. $Q > e^{-G_r^0/RT}$

The quantity $e^{-G_r^0/RT}$ is usually called the thermodynamic equilibrium constant, k_{eq}. The advantage of using this notation is that we can make it appear in the mass action rate by factorising as follows:

$$v = k^+ \left(s_1 s_2 - \frac{p_1 p_2}{k_{eq}} \right)$$

Note about the definition of k_{eq}: In the previous example k_{eq} is dimensionless since it is identified to $e^{-G_r^0/RT}$, and so is the quotient of reaction Q. Let us look at an other example of a reaction where the number of substrates and products are different. If

$$S_1 + S_2 \rightleftharpoons P_1$$

then

$$\Delta G_r = \Delta G_r^0 + RT \log \left(\frac{p_1 c^0}{s_1 s_2} \right)$$

Defining $k_{eq} = e^{-G_r^0/RT}$ would impose the use of c^0 in the mass action rate law. This is possible but often not desirable. Instead, another convention exists for Q and k_{eq}:

$$Q = \frac{p_1}{s_1 s_2} \quad \text{and} \quad k_{eq} = \frac{p_1^{eq}}{s_1^{eq} s_2^{eq}} \tag{2.3}$$

where the "^{eq}" subscript stands for the equilibrium values. This notation is the one used in this thesis. The disadvantage of this notation is that one needs to know the convention for the standard concentration c^0 when evaluating the value of k_{eq} from $\exp(\Delta G_r^0/RT)$ which is dimensionless. The most natural choice is $c^0 = 1\, mmol\, L^{-1}$ because it is closer to the concentrations found in a bacterial cell than $1\ molL^{-1}$.

2.1.3 Michaelis-Menten-Henri

The speed of many chemical reactions can be increased by the introduction of a catalyst in the medium. A catalyst has the property of binding to the substrates but being released at the end of the reaction. In biochemical processes, the role of the catalyst is assumed by dedicated proteins called enzymes. Without enzymes, most biochemical reactions would not occur at a rate that would sustain vital functions. To see how enzymes work, we show in Fig.2.1 a sketch of the transition-state theory [4, 21] representing how the enzyme acts. This diagram illustrates an important feature of many chemical reactions: although they may be favorable in the thermodynamic sense, an initial input of energy is necessary for the reaction to proceed. After that initial boost, the reaction releases an amount of energy larger than the activation energy so that the balance is globaly a release of energy. The function of the enzyme is to lower the activation energy and as a consequence to speed up the reactional processs. Note that the total amount of energy released is the same whether or not the reaction is catalysed.

Enzymatic reactions cannot be modelled by mass action rate laws. An enzyme has a limited catalytic capacity, when all the active sites of all the enzymes are occupied, adding more substrate does not increase the flux of the reaction [35]: the only solution to improve the flux is to add more enzymes [60]. In 1913, Leonor Michaelis and Maud Menten proposed a mathematical model to represent this saturation in a simple case of the irreversible conversion of one substrate into one product [54]:

$$S + E \rightleftharpoons C \rightarrow E + P \tag{2.4}$$

The substrate (S) binds to the enzyme (E) to form a complex (C). The complex is then converted into a product (P) which releases the enzyme. Note that the last step is taken to be irreversible. The kinetics of this reaction are then given by the so called Michaelis-Menten-Henri (MMH) rate law:

$$v(s) = k_{cat} e \frac{s}{K^M + s} \tag{2.5}$$

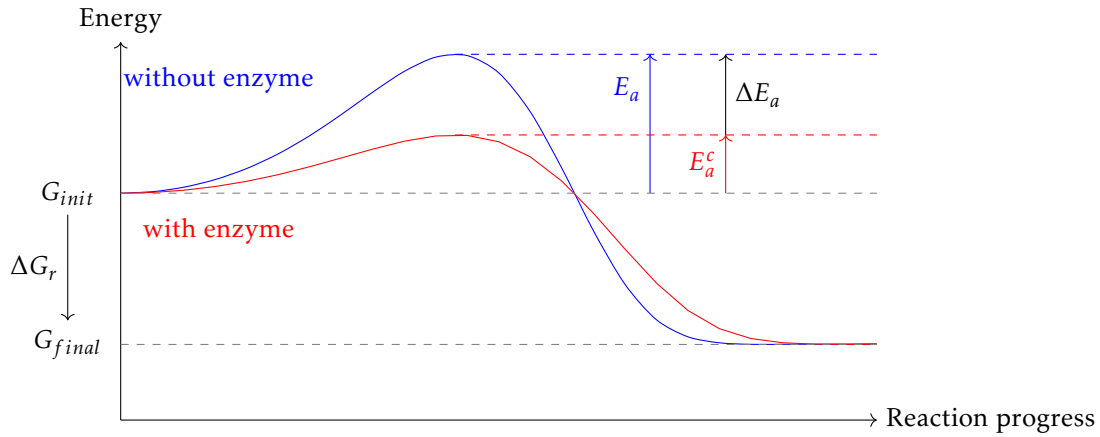


Figure 2.1: Energy of the system as a function of the state of a single molecular reaction event. The reaction releases per mole an energy $|\Delta G_r|$ and each microscopic conversion requires an activation energy E_a to take place, this activation energy is reduced to E_a^c when it is catalysed by an enzyme.

where e and s are respectively the enzyme and substrate concentrations. $V_m = k_{cat}e$ represents the maximum rate for a given concentration of enzymes and K^M, the Michaelis constant, represents the concentration of substrate for which the flux is half of its maximum value. k_{cat} is the catalysis constant which scales with the probability per unit time for a substrate molecule to interact with an enzyme molecule and be transformed into the product molecule. The derivation of such of formula relies on the assumption that the enzymatic complex is at the steady state. For completeness, the derivation is presented in annexe A. In Fig.2.2 we represent the dependence of v on substrate concentration. For small substrate concentrations, the rate increases linearly with s whereas for large concentration the rate saturates, going to a limiting value.

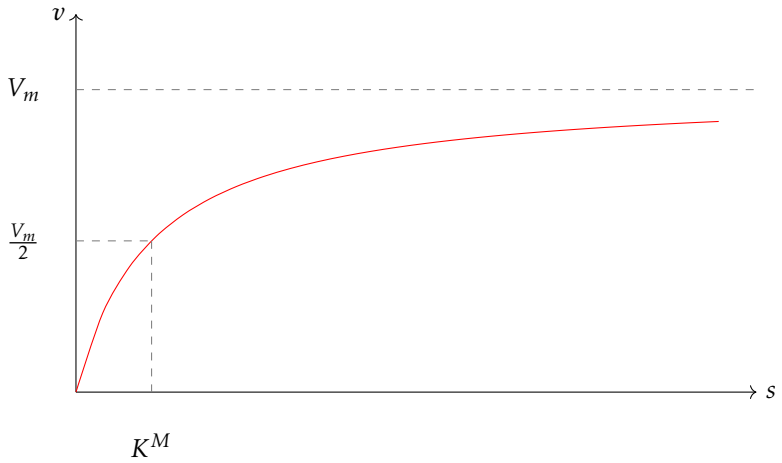


Figure 2.2: Michaelis-Menten reaction rate profile. At low concentrations (s much smaller than K^M), the rate grows linearly with s, but then saturates when the substrate concentration becomes much larger than K^M.

After the foundations of the mass action and the Michaelis-Menten-Henri laws had been laid down, a lot of work investigated various extensions. For instance the MMH rate law lacks reversibility and thus proper thermodynamics, it is reliable only for thermodynamically highly favorable reactions. A

generalised reversible form, known as reversible MMH, is as follows:

$$v(s,p) = \frac{V_m}{K_S} \frac{s - p/k_{eq}}{1 + \frac{s}{K_S^M} + \frac{p}{K_P^M}} \tag{2.6}$$

This equation is derived by allowing the second step of Eq. 2.4 to be reversible. The substrate and the product have their own Michaelis constants. It has the advantage of taking the thermodynamics into account for a one substrate-one product reaction.

2.1.4 Rates for higher order reactions and the King-Altman method

In the previous paragraph, the assumption that the intermediate enzymatic complex was at a steady state allowed one to determine the rate law for one substrate and one product. However, for more substrates or more products, the manual derivation of the overall rate law is no longer feasible. Instead, one resorts to an algorithm based on Cleland's graphs. A Cleland's graph represents an enzymatic reaction via a graph where each node represents an enzymatic complex and where the edges, linking the states of the enzyme, represents the binding or unbinding of a metabolite to the complex, these edges thus representing elementary steps. This nomenclature gives information about the underlying mechanism of the (complex) reaction. Let us illustrate this approach with the example of a reaction with two substrates and two products:

$$S_1 + S_2 \overset{(E)}{\rightleftharpoons} P_1 + P_2 \tag{2.7}$$

Different scenarios exist depending on which substrate binds to the enzyme first and which product leaves the enzyme first. (i) One can have an *ordered bi-bi* mechanism where the two substrates bind successively and then the two products are released successively; four associated orders are possible. (ii) One may have events following a *ping pong* pattern where the first substrate binds to the enzyme and it is converted into the first product which leaves the *enzyme* in an activated state, ready to react with the second substrate to produce the second product. Here again, there are four ordering possibilities depending on which of the two substrates binds first and which product is released first. (iii) A last *random bi-bi* mechanism can arise where the order in which the metabolites join the reaction does not matter. In the last case, the two substrates have to be recruited before the first substrate is released otherwise it would lead to two successive unimolecular reactions. The three mechanisms are represented in Fig. 2.3.

In Fig. 2.3, the elementary reaction decomposition produces a set of reactions acting according to a mass action rate law, each of which has its own kinetic constants. Assuming the steady state for the enzymatic complexes or the individual steps, a set of relations can be written between the complexes' concentrations, the metabolic concentrations, and the total concentration of enzyme (a derivation is presented in annexe A). Solving the system of equations gives the rate of the total reaction. However the solution can be hard to determine, even numerically, so it is common to use a method proposed by King and Altman [48]. With this technique, one can find the rate law for any reaction once the mechanism is specified. It also can be used to model cases where an activator or inhibitor controls a reaction.

For reactions of the type Eq. 2.6 but generalized to multiple substrates and/or multiple products, the Cleland procedure leads to a mass action numerator and a saturation term in the denominator which is a polynomial of the metabolic concentrations:

$$v(S_1,S_2,P_1,P_2) = V_m \frac{s_1 s_2 - p_1 p_2/keq}{\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{m=0}^{1}\sum_{n=0}^{1} K^{(i,j,m,n)} s_1^i s_2^j p_1^m p_2^n} \tag{2.8}$$

Such a rate law requires a very high number of parameters. Despite its faithfulness to the actual mechanism taking place, its complexity makes it quite unpopular: in practice, it is often rejected in favor of simpler representations for which the parameter estimation has a higher chance to succeed.

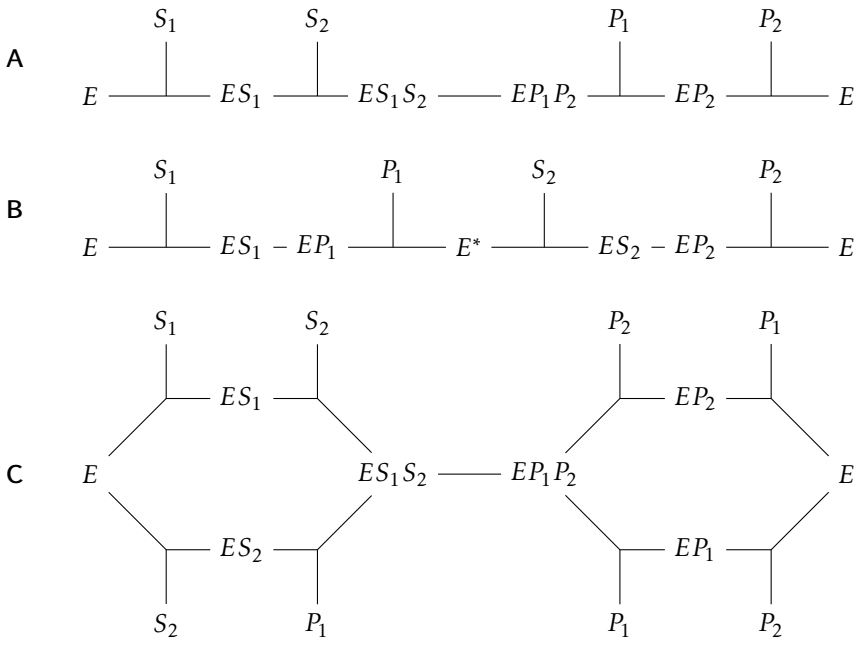


Figure 2.3: The three mechanisms represented by Cleland's graphs in which two substrates leading to two products. **A** represents an ordered bi-bi mechanism, **B** a ping pong mechanism and **C** a random bi-bi mechanism. Each segment represent a reactional step.

2.1.5 The Lin-Log formalism

The lin-log kinetics correspond to another formalism for specifying kinetic rate laws. In the example of Eq. 2.7, the rate would be

$$v = \alpha e(1 + l_1 \log(s_1) + l_2 \log(s_2) + h_1 \log(p_1) + h_2 \log(p_2)) \tag{2.9}$$

This form was introduced by Visser and Heijnen [77]. It is not derived from mechanistic laws; rather, it is motivated by the relation between the rate of a reaction and the thermodynamic driving force, in our case ΔG_r. Inspired by conduction processes, the flux is taken to be proportional to the driving force [59] which is nothing else than a linear combination of the logarithms of the concentrations (Eq. 2.2). Although the formalism is both approximate and a bit ad-hoc, it has gained in popularity due to its relatively small number of parameters. Furthermore, one can reformulate the rate law so that the elasticity parameters relative to a reference state S^0 appear explicitly:

$$v = J^0 \frac{e}{e^0}\left(1 + \epsilon^v_{S_1} \log\left(\frac{s_1}{s_1^0}\right) + \epsilon^v_{S_2} \log\left(\frac{s_2}{s_2^0}\right) + \epsilon^v_{P_1} \log\left(\frac{p_1}{p_1^0}\right) + \epsilon^v_{P_2} \log\left(\frac{p_2}{p_2^0}\right)\right) \tag{2.10}$$

A term having the "^{0}" superscript is associated with the reference state S^0 under consideration, so for instance J^0 and e^0 are respectively the corresponding flux and enzyme concentration. The $\epsilon^v_X = \frac{x^0}{J^0}\frac{d,v}{dx}|_{x_0}$ coefficient is the scaled elasticity of the metabolite X and accounts for how sensitive the flux v is to the metabolite's concentration. Note that the formula Eq. 2.10 is exact in state S^0, a nice feature which ensures that the lin-log formalism can be used near the reference state with good accuracy. Furthermore, in the limit where one substrate or product arises in infinitesimal concentrations, the flux will be oriented towards the production of that metabolite as it should. Lastly, in contrast to Michaelis-Menten-Henry, the flux does not saturate as one of the concentrations diverges, though its *rate of increase with concentration* does go to zero.

2.1.6 The convenience kinetics formalism

For enzymatic reactions, we have seen two types of kinetic rate laws so far. The first relies on the knowledge of the exact mechanism for the reaction steps whereas the second only needs the list of metabolites that are involved in the reaction, approximating the flux from the elasticities and a known reference state. The convenience kinetics formalism [51] introduces a form for the rate laws as an alternative for coping with reactions for which the mechanism is unknown. Compared to the lin-log formalism, it has the advantage that it imposes zero flux when the system is at equilibrium. It does so by enforcing thermodynamics via the equilibrium constant. Possible saturation effects in the substrates or products are encorporated by generalizing the functional form used in Michaelis-Menten-Henry. Specifically, for any set of substrate $\{S_i\}_{i=1\ldots N}$ and products $\{P_j\}_{j=1\ldots M}$, the kinetic law specified by convenience kinetics is

$$v(\{s_i\},\{p_j\}) = \frac{V_m}{\prod_{i=1}^{N} K_{S_i}} \frac{\prod_{i=1}^{N} s_i - \prod_{j=1}^{M} p_j/k_{eq}}{\prod_{i=1}^{N}\left(1+\frac{s_i}{K_{S_i}}\right)+\prod_{j=1}^{M}\left(1+\frac{p_j}{K_{P_j}}\right)} \tag{2.11}$$

where s_i and p_j are respectively the concentration of S_i and P_j. V_m is the maximum rate constant, proportional to enzyme concentration. The K_X constant is the dissociation constant for the metabolite X and k_{eq} is the thermodynamics equilibrium constant. The dissociation constant embodies the concentration over which the corresponding metabolite contributes significantly to the saturation effect. Interestingly for one substrate and one metabolite, the equation Eq. 2.11 is of the same form as Eq. 2.6 where the dissociation constants play the role of the Michaelis constants. Although it does not contain regulation effects, the kinetic law can easily be modified to account for metabolic activations or inhibitions.

2.1.7 Rate laws chosen for this thesis

The rate law specified by convenience kinetics is ideal in the context of developping a systematic methodology for the construction of metabolic kinetic models since convenience kinetics do not require knowing the precise underlying mechanism for each reaction. Note that even if the mechanism *is* known, convenience kinetics provide a good tradeoff between a small number of parameters and an accurate description of the reaction. It is also more appropriate than a lin-log rate law when one wants to explore a network for several different steady states and thermodynamic validity is important. For modeling activation or inhibition, it is common practice to change the convenience kinetics rate law by multiplying the expression by a term ranging between 0, for total inhibition, to 1, for total activation:

$$\frac{a}{K_A + a} \qquad \frac{K_I}{K_I + i} \tag{2.12}$$

In these factors, the activator and inhibitor concentrations are denoted by a and i. A characteristic concentration K_A accounts for the concentration scale where the activator (resp. inhibitor) begins to have a significant effect. With this framework for rate laws in our modelling, we are thus able to take into account

- the thermodynamics of the reaction
- a saturating effect on the flux by the substrates and products
- simple regulatory or allosteric phenomena arising from metabolites or other effectors

However the following phenomena are not modelled in our work:

- precise mechanisms for the enzyme kinetics
- effects of the ionic environment (pH, metallic ions)
- higher levels of regulation such as transcriptional inhibition or post-transnational modifications

2.2 Determining parameters of the rate equations

2.2.1 Experimental techniques for measuring k_{eq}

The equilibrium constants are probably the quantities for which the information is the easiest to get since the associated values can be infered directly from experiments or from known reactional Gibbs' energies of reaction using $k_{eq} = \exp\left(-\frac{\Delta G_r}{RT}\right)$. ΔG_r is independent on the catalyzing enzyme, which means that the value does not change accross different organisms. Of course ΔG_r is dependent on the concentrations of the reactants and on the temperature. For this reason, it is common to store the ΔG_r^0 values in tables. This quantity stands for the reactional Gibbs energy for standard conditions of temperature (T=293.15 K), concentration (1 $mol\,L^{-1}$), and pressure (1013.25 hPa). The reactional energies under different conditions are easily found using Eq. 2.2.

Reactional Gibbs energies are potentials: they do not depend on the path followed by the system to go from a state before the reaction to a state after the reaction. This is very convenient because one may not know the energy characterizing a given reaction but one can still derive a succession of steps for which the energy transition is known and infer the unknown energy. To do so, molecules are labelled with standard formation energies ΔG_f^0 that define the energy needed to build the molecule from pure compounds in their standard form. By convention, the energy of formation of pure compounds under their standard form is set to 0. For example, water is formed by the reaction

$$H_2^{(g)} + \frac{1}{2} O_2^{(g)} \rightarrow H_2O$$

The energy released by this reaction is $-158.9\ kJ\,mol^{-1}$, because the energy of formation of the two pure compounds $H_2^{(g)}$ and $O_2^{(g)}$ is 0, the energy of formation of water is $-158.9\ kJ\,mol^{-1}$. Now we can extend our example to the reaction

$$\begin{aligned} ATP + H_2O \rightarrow ADP + P_i \qquad & (R1) \\ \Delta G_r^0(R1) \quad = \quad & \Delta G_f^0(P_i) + \Delta G_f^0(ADP) - \Delta G_f^0(H_2O) - \Delta G_f^0(ATP) \end{aligned}$$

This reaction is almost irreversible and there is no hope to measure the apparent k_{eq}. However the energies of two other reactions have been measured:

$$\begin{aligned} ATP + Glc \rightleftharpoons ADP + G6P \qquad & (R2) \\ \Delta G_r^0(R2) = -25\ kJ\,mol^{-1} \quad = \quad & \Delta G_f^0(G6P) + \Delta G_f^0(ADP) - \Delta G_f^0(ATP) - \Delta G_f^0(ATP) \\ G6P + H_20 \rightleftharpoons Glc + P_i \qquad & (R3) \\ \Delta G_r^0(R3) = -12\ kJ\,mol^{-1} \quad = \quad & \Delta G_f^0(P_i) + \Delta G_f^0(Glc) - \Delta G_f^0(H_20) - \Delta G_f^0(G6P) \end{aligned}$$

Combining the formation energies of the molecules involved in $R2$ and $R3$ allows us to recover $\Delta G_r^0(R1) = \Delta G_r^0(R2) + \Delta G_r^0(R3) = -37\ kJ\,mol^{-1}$.

For oxidation–reduction (redox) reactions, the reactional Gibbs is obtained from the electrical potential of a Galvanic cell

$$\begin{aligned} O_1 + R_2 \rightleftharpoons O_2 + R_1 \quad & \text{redox reaction} & E &= E_1 - E_2 \\ O_1 + n_1 e^- \rightleftharpoons R_1 \quad & \text{half-cell 1} & E_1 &= E_1^0 + \frac{RT}{n_1 F} \log(O_1/R_1) \\ O_2 + n_2 e^- \rightleftharpoons R_2 \quad & \text{half-cell 2} & E_2 &= E_2^0 + \frac{RT}{n_2 F} \log(O_2/R_2) \end{aligned} \tag{2.13}$$

$$\Delta G_r = -n\,F\,E \tag{2.14}$$

with n being the number of exchanged electrons and $F = 96\,485\ C\,mol^{-1}$ being the Faraday constant. Indeed, the electrical potential of each half cell can be measured with a simple voltmeter and is defined as in Eq. 2.13. Identification of the terms of Eq. 2.2 with the terms of Eq. 2.13 results in the expression for ΔG_r as defined in Eq. 2.14. This methods provides a large number of energies of formation.

A first table using observed equilibrium constants and combinatorial rules as illustrated in the former examples was first published 1957 [49] for about a hundred formation energies. This table was then followed by several extensions [29,73] and the most up-to-date version was published by Alberty in 2006 [3].

The energy of a reaction may be affected by a variation in the pH or in the ionic strength which is something that has not been discussed so far. Formulas exist to account for these effects and are discussed in Alberty's work [2]. Another important point that has not been mentioned yet is that the formation energy for some cofactors is set to 0 in many database. The reason for this is that reactions involving cofactors use them as conjugated pairs: NAD/NADH, ADP/ATP, ... In such conditions, only the energy of transition between the conjugated cofactors is required. For example, Alberty's table associates 0 to the formation energy of adenosine and NAD.

2.2.2 Using theory to calculate k_{eq}

Using experimental data for apparent equilibrium constants and combining them together allows one to determine quite a number of Gibbs formation energies. The most recent works provide about 400 metabolic reactions that have been measured or for which the Gibbs energy has been calculated. Unfortunately this is far from covering the totality of known metabolic reactions, for example the *E. coli* bacterium on its own (and which is a quite "simple" organism) has about 720 reactions. Because of this limitation, further approaches are necessary to estimate reaction energies.

The first method, proposed by Burton and Buss [9], consists in assuming that a chemical bond between two types of atoms always has the same energy, regardless of the other atoms in the considered molecule. The methods was then refined by considering the atoms in a close neighborhood of 3-5Å, called a group, rather than focusing on a pair of atoms at a time. The reactional Gibbs energy is calculated by adding up the energies of the bonds (resp. groups) formed in a reaction. These methods have extended the number of available Gibbs reactional energies significantly but at the price of a loss in precision, the error reaching 10 $kJ\,mol^{-1}$. Furthermore, they cannot be used with experimental values without a minimum amount of care because there is the risk of causing a violation of the first law of thermodynamics.

Noor et al. proposed a way to unify the bond and group contributions in a theory called component contributions [58]. The approach prioritizes experimental reactant contributions (energy of formation) over group contributions. We will summarize here the main concepts of their technique but a more extensive description can be found in the original paper. For a set $\Delta_r G^0_{obs}$ of measured reactional energies, one may infer formation energies for the reactants using the pseudo-inverse of the transposed stoichiometric matrix $S \in \mathbb{R}^{m\times n}$:

$$\Delta_f G^0_{rc} = (S^T)^+ \Delta_r G^0_{obs}$$

Note that when the model is over-constrained by $\Delta_r G^0_{obs}$, the excess of information is handled by using the null space of $(S^T)^+$. This solves the problem of potential contradictions, due for instance to measurement errors. Then, new reactional energies consistent with the thermodynamics can be recalculated:

$$\Delta_r G^0_{rc} = S^T (S^T)^+ \Delta_r G^0_{obs}$$

An estimate of the error on the $\Delta_r G^0_{rc}$ is given by $\Delta_r G^0_{rc} - \Delta_r G^0_{obs}$.

For the group contributions, it is assumed that each reactant's energy of formation is a linear combination of the energetic contributions from the groups it contains, *i.e.*, $\Delta_f G^0 = \mathcal{G}\Delta_g G^0$ with $\mathcal{G} \in \mathbb{R}^{m\times g}$. Using a similar relation as for the reactant energy of formation, the group contributions are given by

$$\Delta_g G^0_{gc} = (S^T \mathcal{G})^+ \Delta_r G^0_{obs}$$

Again the thermodynamic inconsistencies have been removed. The underlying idea of the component contributions method is to decompose a reaction into a part that involves measurable reactions and another part that does not but for which one can use group components contributions. The stoichiometry of any reaction decomposes as $x = x_R + x_N$ where x_R is in the range of S (combination of measured reactions) and x_N is in the null space of S^T (so it cannot be decomposed into a linear combination of the

stoichiometry of the measured reactions). The choice made in [58] considers the orthogonal projection $P_{\mathcal{R}(S)}$ of x onto the range of S and $P_{\mathcal{N}(S^T)}$ for the projection of x on the null space of S^T. In this way, a reaction with a stoichiometric decomposition x along the m reactants of the network has a standard reactional Gibbs energy

$$\Delta_r G^0_{cc,x} = x^T (P_{\mathcal{R}(S)} (S^T)^+ + P_{\mathcal{N}(S^T)} \mathcal{G} (S^T \mathcal{G})^+) \Delta_r G^0_{obs}$$

In practice, this method is very convenient because it provides us with reactional Gibbs energies that are thermodynamically consistent. Furthermore, compared to the approach using only formation energies, the components contributions method extends the list of reactions for which the energy can be computed. Note that it is important to keep in mind that when no measured reactions provide information for the formation of a group, all the reactions containing this group will still have an unknown energy of formation.

2.2.3 Time series to measure $\mathbf{k_{cat}}$ and $\mathbf{K^M}$

We now consider the determination of the parameters k_{cat} and K^M. These are typically evaluated from time course data. For an arbitrary time course, data are not so easy to exploit. Instead, one focuses on the reaction velocity at early times, before any product has had time to accumulate. Let $v_0(s_0)$ be the initial rate of the reaction when the initial concentration of substrate is s_0. (Naturally this can be generalized to the case of more than one substrate.) If one can measure $v_0(s_0)$ for several controlled and known values of s_0 and assuming known the concentration of enzymes, the values of $\mathbf{k_{cat}}$ and $\mathbf{K^M}$ can be determined. Typically one would like to measure several $v_0(s_0)$ for each reactant involved in the reaction, this way the K_m associated to every reactant can be obtained. As for k_{cat}, it is in fact obtained from the maximum rate V_m, which is easier to obtain, by dividing by the concentration of enzyme, e, if it is known. In practice, v_0 is obtained by stopping the reactions (e.g. by diluting the medium) after some time (during which the products have not accumulated too much), measuring the concentration of substrate or product, and then correcting for dilution, obtaining the amount of consumed substrate and thus the rate of the reaction.

Let us illustrate the inference of V_m and $K^M_{(F6P)}$ for the reaction catalysed by Pgi ($F6P \rightleftharpoons G6P$). To measure $K^M_{(F6P)}$, the initial concentration of G6P is kept as low as possible and the reaction is started with different initial concentrations $[F6P]_0$, after a few minutes the new concentration is measured. Using the concentration difference and the time during which the reaction proceeded gives us the initial rate for one $[F6P]_0$:

$$v_0 = V_m \frac{[F6P]_0}{K^M_{(F6P)} + [F6P]_0} \tag{2.15}$$

Experimental values are displayed in Fig. 2.4. Fitting Eq. 2.15 to the data provides the kinetic parameters of interest. One should keep in mind that the parameter values depend on the model used for the fit which means that the selected model has to be chosen carefully. In particular the reaction Pgi is inhibited by PEP which was not taken into account here since the in vitro measurement has non. However, fitting in vivo data with Eq. 2.15 would have led to a different apparent V_m^{app}. The Pgi reaction will be revisited in this thesis. By standard convention, its forward direction goes from G6P to F6P. Consequently, the measured V_m as described contains in fact the backward $k^-_{cat} = V_m/e$. The forward k^+_{cat} is obtained from

$$k^+_{cat} = \frac{V_m k_{eq}}{e}$$

The next step would be to use the same method with data $v_0([G6P]_0)$ to get $K^M_{(G6P)}$. Given V_m, $K^M_{(G6P)}$ and $K^M_{(F6P)}$, the reversible MMH model for Pgi is then completely determined:

$$v([G6P],[G6P],e) = V_m e \frac{[G6P] - [F6P]/k_{eq}}{1 + [G6P]/K^M_{(G6P)} + [F6P]/K^M_{(F6P)}}$$

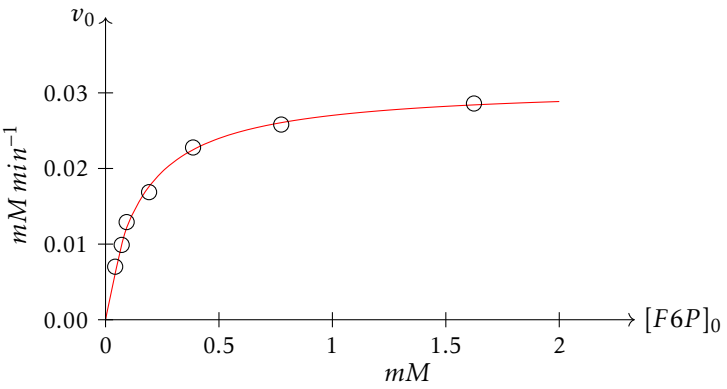


Figure 2.4: Experimental values [27] for the initial reaction rate are shown (circles) for different values of intial F6P concentrations. The best fit with a Michaelis-Menten-Henri behaviour is plotted in red. The parameters are $V_m = 0.031\ mM\,min^{-1}$ and $K^M_{(F6P)} = 0.15\ mM$.

2.3 Reaction networks in central carbon metabolism

The central carbon metabolism (CCM) is a very important part of the global metabolism due to its key role in extracting energy and carbon precursors from external compounds. It is present with some variations in almost all organisms, and it has been particularly well studied in *E. coli* . The CCM produces energy and the essential precursor metabolites required amongst others for *de novo* synthesis of amino acids, the building blocks of proteins. Here will describe the three pathways included in the CCM: the glycolysis pathway, the pentose phosphate pathway(PPP) and the tricarboxylic acid (TCA) cycle (also known as Krebs cycle). In this presentation, I will put a strong focus on the case of *E. coli* .

2.3.1 The glycolysis pathway

It is possible that glycolysis played an important role in the primary anaerobic organisms which produced energy in environments without oxygen. It is still present in contemporary organisms as the backbone of the CCM. Its name comes from its ability to break-down glucose, and traditionally the glycolysis pathway s considered to end with pyruvate. The glycolysis process, from glucose to pyruvate, produces energy for the organism under the form of the high-energy compounds ATP and NADH. In its most common version, the Embden-Meyerhof-Parnas pathway (EMP), glycolysis can be divided into two stages, a preparatory phase that needs two molecules of ATP in order to absorb glucose and to transform the fructose-6-phosphate(F6P) into fructose-1,6-biphosphate (FbP), and a "pay-off" phase where, under optimal conditions, a maximum of four molecules of ATP are released by the reaction from 1,3-biphosphoglycerate to 3-phosphoglycerate and the reaction from PEP to pyruvate. Each of thes! e reactions releases one ATP, but there is a factor of two coming from the splitting of a 6-carbon compound, FbP, into two 3-carbon compounds, dihydroxyacetone phosphate (DHAP) and glyceraldehyde-3-phosphate (GAP). When each of these compounds takes part in the pay-off phase, the flux for this phase is twice the flux of the preparatory phase, then the total yield of the glycolysis is two ATP molecules per incoming glucose molecule.

E. coli possesses an alternative pathway that also serves as glycolysis: the Entner-Doudoroff (ED) pathway. Its ATP yield per glucose molecule is one (instead of two for the EMP pathway) so one may question its usefulness. However, it has been demonstrated that the ED pathway requires less enzyme, and so the so-called "protein cost" may be less [24]. In a medium which is poor in nitrogen sources and where all the amino acids have to be synthesized de novo, the ED pathway may be thus preferred over the EMP, in spite of its lower yield in ATP.

A second function of the glycolysis pathway is to produce precursors for the other pathways of the CCM, precursors that often are necessary for biomass production. Glucose-6-phosphate (G6P), GAP and F6P and pyruvate (Pyr) serve as precursors for the PPP and uvate (uvate) serves for the TCA. PEP and

Pyruvate are directly involved in the synthesis of amino acids. A more visual description of glycolysis is represented in red in Fig. 2.5.

2.3.2 The pentose phosphate pathway

This pathway is parallel to the initiation steps of the EMP glycolysis and has many connections with it (G6P, F6P, GAP). It also shares its two first steps with the ED pathway. It can thus function to bifurcate away from the glycolytic pathway. As for glycolysis, its origin is quite ancient in evolution. In so-called less primitive organisms, the reactions in the PPP are catalysed by enzymes, but in some archae for instance, catalysis is performed by metallic ions.

The first phase of the PPP pathway is oxydative; the conversion of two molecules of NADP to NADPH is responsible for an important part of the cell's oxydative power. The second phase is non-oxydative and produces the important biochemical precursors erythrose-4-phosphate (E4P) and ribose-5-phosphate (R5P). These metabolites are used for biosynthesis of amino acids while R5P serves in nucleotide production. The PPP is coloured in green in Fig. 2.5.

2.3.3 The tricarboxylic acid cycle

Glycolysis is able to produce two ATP molecules from one glucose molecule. This ratio accounts only for a small part of the potential energy contained in glucose. Indeed, in favorable conditions, far more energy can be extracted by the TCA cycle which processes pyruvate. Under aerobic conditions, the 3-carbon pyruvate is oxydised into three molecules of carbon dioxide. The first oxydation converts pyruvate to the 2-carbon acetyl-coa (AcCoA) which feeds the TCA cycle. That cycle consists of eight reactions that start with the binding of AcCoA to oxaloacetate (OAA) to produce citrate (Cit). A molecule of cytrate is then oxydised OAA, releasing two molecules of CO_2. The cycle is catalysed by a set of enzymes associated with the cell membrane among which are the ATPases that can use the proton gradient between the cytoplasm and the periplasm to produce ATP from ADP with a ratio of approximately $4H^+$:1ATP [70]. Considering that the oxydative power of! NADH can translocate one $H^+$and the oxydised form of coenzyme Q(QH_2) 6 proton [65], the TCA cycle is able to produce a maximum of 25 ATP molecules from one glucose molecule as shown in Tab. 2.1.

Reaction	**ATP equivalent**
Pdh	1 NADH $\rightarrow$ 10 $H^+\rightarrow$ 2.5 ATP
Icdh	1 NADH $\rightarrow$ 10 $H^+\rightarrow$ 2.5 ATP
Kgdh	1 NADH $\rightarrow$ 10 $H^+\rightarrow$ 2.5 ATP
Stk	1 ATP
Sdh	1 $QH_2 \rightarrow$ 6 $H^+\rightarrow$ 1.5 ATP
Mdh	1 NADH $\rightarrow$ 10 $H^+\rightarrow$ 2.5 ATP
Total (1 Pyr)	12.5
Total (2 Pyr)	25

Table 2.1: Number of ATP equivalents produced by the TCA reactions. The names of the reactions are those used in Fig. 2.5. The conversion ratios used are 1NADH:10H^+, 1QH_2:6$H^+$and $4H^+$:1ATP.

The energy production is not the only interesting aspect of the TCA, many important biomass precursors are also produced. In particular AcCoA, OAA, SucCoA, and αKg are produced and these are used for amino acid production. The TCA is coloured in blue in Fig. 2.5.

2.3.4 Acetate secretion

Two additional reactions are responsible for the secretion of acetate: one converts AcCoA to acetate and one allows the excretion of acetate, transporting it from within the cell to the exterior. The secretion of acetate can be considered as a procedure to maintain flux in the glycolysis pathway (with its production of energy) by discarding carbon atoms when the TCA is saturated and cannot absorb the necessary flux.

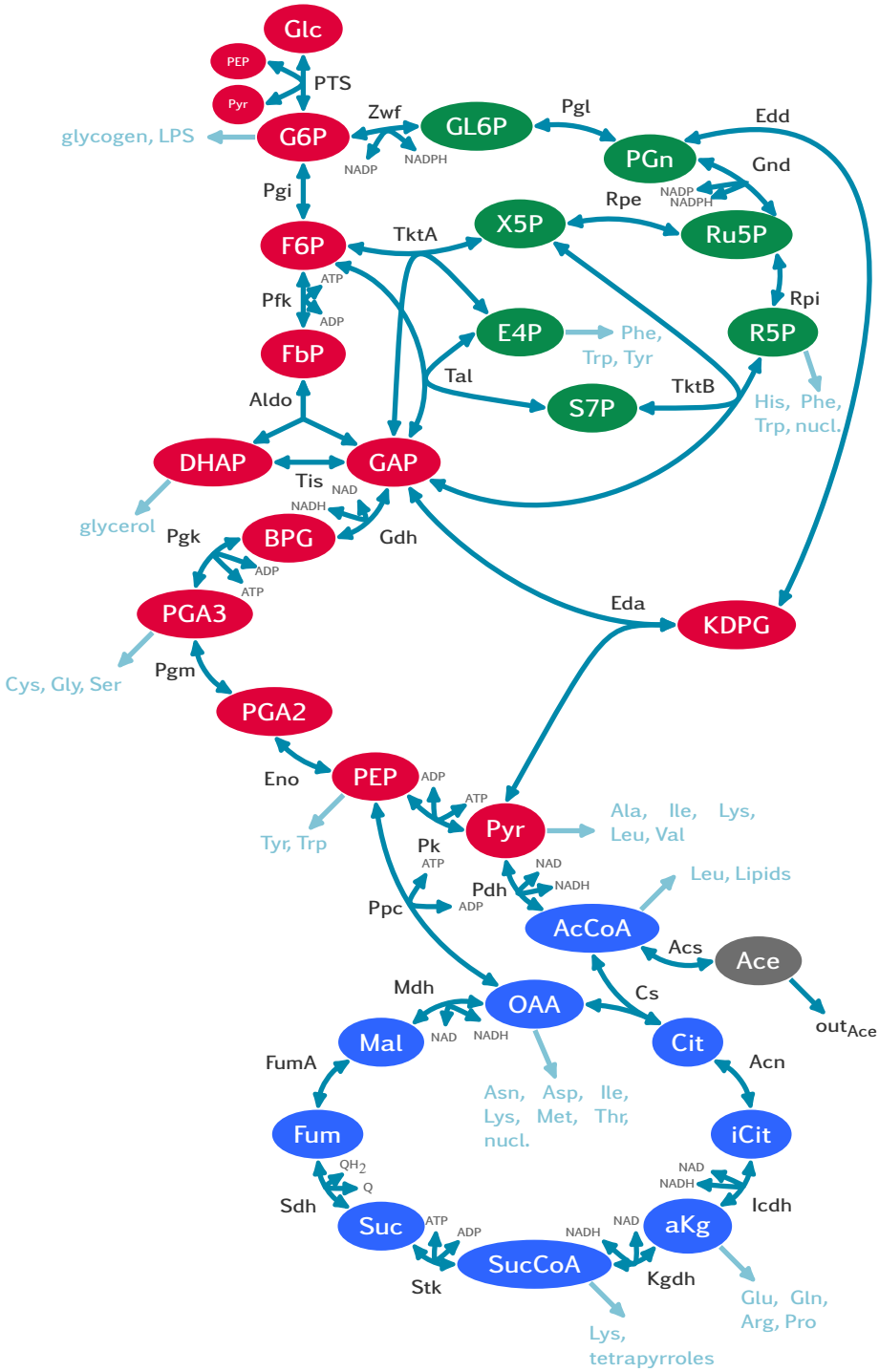


Figure 2.5: Reactions of the central carbon metabolism. The CCM contains the three pathways: glycolysis, pentose phosphate and tricarboxylic acid cycle, coloured respectively in red, green and blue. Enzyme-catalysed reactions are displayed using bidirectional arrows with the cofactor taking part in the reaction. The contribution of the various precursors to other pathways are also presented in light blue.

2.4 Determining systemic properties of metabolic networks

A recurring difficulty arising in the construction of kinetic models is the lack of data for adjusting such models. In the previous sections I covered methods for constraining kinetic parameters arising in rate laws; now I will focus on approaches to quantify enzymes, metabolite concentrations and fluxes.

2.4.1 Measurements of concentrations of enzymes

Within a few years, it should be possible to determine average concentrations of all enzymes in an organism. At present though, absolute quantification is difficult and limited to just a handful of enzymes at a time. Relative quantification is simpler but still suffers from systematic biases. In the long run, it may be possible to get away from population averages and determine single cell protein concentrations but currently quantification's in single cells are really feasible only for RNA and DNA. In the following, I present an overview of a number of techniques for identifying enzymes and for quantifying their abundance in populations.

Two-dimensional gel electrophoresis: This method exploits physical properties that allow one to distinguish proteins according to their migration capacity. When exposed to an electric field, ions migrate: positively charged ions will drift toward the cathode whereas negatively charged ions will drift toward the anode. In solution, proteins are ionized and in fact their residues can be found under different ionized forms as in Eq. 2.16. Depending on the pH conditions (concentration of H^+ ions) in the medium, the global charge of a protein will vary. There exist a particular pH, the isoelectric point, for which the molecule is neutral and thus is not affected by an electric field. It then does not drift away from such a point and instead will accumulate there (the isoelectric point is stable). Since proteins have different chemical compositions, the isoelectric point tends to be like a finger-print, specific to one protein. This point can be r! ealized by placing a gel between a cathode and an anode to create an electric current while at the same time having two different buffers at each end of the gel to produce a pH gradient. In such a gel, the (charged) proteins will migrate until they reach their isoelectric point where they are no longer charged. An initial blob of protein will thus separated into a set of spots, each formed typically of a single protein species.

$$\begin{aligned} R-COOH &\leftrightarrow R-COO^- + H^+ \\ R-NH2 + H^+ &\leftrightarrow R-NH3^+ \end{aligned} \tag{2.16}$$

To improve the resolution i.e., to better separate the spots, one usually resorts to 2-dimensional gel electrophoresis. The reason is that it may happen that by accident two proteins have a very close isoelectric point so that their two spots in fact coincide or strongly overlap. To separate these proteins, a second electrophoresis is applied in another dimention of the gel (typically, the gel is turned by an angle of 90^o). After the first migration, the proteins are denatured to unfold them and they are treated with sodium dodecyl sulfate (SDS) which will bind the protein along its (unfolded) length; that way the charge of the protein is mainly due to the negatively charged SDS. Applying the rotated electric field will then separate the proteins along this new dimension according to the number of SDS bound to them. That number is proportional to the size of the protein, i.e., its molecular weight, so this 2-dimensional gel approach allows one to separate proteins accordi! ng to both isoelectric point and molecular size.

In many studies, a 1-dimensional electrophoresis is sufficient to separate the enzymes of interest. The 2-dimensional version shows its value when a better resolution is required. During the first migration, proteins that have naturally a strong interaction (have formed a complex) will migrate together but during the second migration they will separate. Spots present on the same line of the second migration may thus indicate the presence of such complexes, i.e., protein-protein interactions. After the electrophoresis is finished, the resulting spots are analyzed either manually or with imaging software to infer the proportion of the different proteins between different samples, obtained for instance under different physiological conditions. To identify the enzyme associated with a spot, ab-initio methods can be used (based on the isoelectric values or mass spectrometry as we will discuss soon). But once the

mapping between spots and proteins has been made in one gel, the map! ping can be used in further experiments without the need for new protein identification procedures.

Liquid Chromatography / Mass Spectrometry (LC/MS): A spectrometer is a device capable of separating ions according to their mass/charge ratio. It contains a region in which an ion is accelerated (using electric or magnetic fields); the value of the mass/charge ratio is then inferred either by the time of flight of by the deflection of the trajectory produced by a magnetic field. Indeed, if a charged particle is subject to an electric or magnetic field, its kinetics depends only on that ratio, not on mass and charge separately. Specifically, the particle's acceleration is given by the Lorentz force divided by that same ratio:

$$\vec{a} = z\frac{\vec{E} + \vec{v} \times \vec{B}}{m\,e} \tag{2.17}$$

where $\vec{a}$ is the particle's acceleration, m its mass, z its charge number, $e = 1.602\,10^{-19}C$ is the elementary charge, $\vec{v}$ the ion's velocity, $\vec{E}$ the electric field, $\vec{B}$ the magnetic field and $\times$ the vector cross product.

There are multiple technologies for spectrometers, and they can roughly be classified into two types.

- The first type relies on the time of flight. The ions are exposed to an electric field and are accelerated proportionally to the value of z/m. The detector measures this time of flight for the ion; the smaller the m/z ratio, the shorter the time of flight.
- The second type includes a magnet after the electric field, thereby curving the ion's trajectory with a force proportional to its velocity. The detector measures the deviation produced by the magnetic field to infer m/z.

The values of m/z for large biomolecules can be quite similar, making it impossible to distinguish different proteins using only that ratio with today's resolution in mass spectrometry. To overcome this difficulty, a first step of digestion splits the proteins into smaller peptides. These peptidic m/z values are better resolved in mass spectrometry and so can be used to identify the peptide content of the sample. The first LC/MS techniques thus required using a single spot of a 2-dimensional gel to produce a sample containing a priori a single protein species. Given this pure sample, the proteins would be digested e.g. by trypsin, leading to a sample of different peptides characteristic of the protein of interest. These peptides were separated by liquid chromatography, a process exploiting the different speed of migration of each peptide. The different constituents exit the LC at different times and are then sent to the mass spectrometer for identification. Given a set o! f such identified peptides, it is often possible to unambiguously determine the protein in the sample, either by ab initio approaches or by comparing the set and intensities of peptides found to results tabulated in databases.

Extracts from cells contain mixtures of proteins and other molecules, justifying the need to use 2-dimensional gel electrophoresis to create samples with a single species of proteins. However, the tedious procedures required for such gels can now be avoided with the technique of so-called MS/MS. In that more recent approach, all proteins in the sample are digested and pushed into the LC. The MS/MS provides a two-step process in which peptides are further broken down, allowing each peptide to be identified via its characteristic m/z spectrum in the second MS. Such an identification can be done de novo or by comparing to theoretical or experimental m/z spectra stored in a database. The list of peptides (and to some extent their abundances via the relative intensities of the peaks) can then be used to reconstruct the complex protein content in the sample. This last step requires computations to disentangle the contributions of all peptides because different proteins can sha! re identical peptides.

Quantitative mass spectrometry: The main disadvantage of the basic LCMS and MS/MS techniques is that they give only relative proportions of peptides and even those are plagued by biases because of the protein digestion phase. A trick to obtain the absolute quantity of a peptide is to add a known amount of its labelled version to the sample. The labeling can be performed by a stable heavy or light isotope. Once mixed into the sample, the labelled and unlabeled peptides migrate with the same speed into the chromatography column. However the mass difference due to the isotope will shift sightly the two signals in the m/z spectrum. The intensity ratio between the signals tells us about the relative

amount of the two peptides. Since the labelled peptide can be introduced in a controlled amount, the relative abundances allow one to determine the absolute amount of the unlabelled peptide as promised.

2.4.2 Measurements of concentrations of metabolites

Another quantity to characterize in order to describe a cell's metabolism is the concentration of the different metabolites. These are small molecules compared to the enzymes and they rarely exceed ten carbons. The procedure to detect their presence is quite simple since they can be ionised and identified using the LC/MS technology just as for peptides. But the approach also allows one to determine abundances from the amplitudes of the signal - ie the area under a peak. That was not the case for proteins because strong biases are introduced the digestion phase for going from proteins to peptides. For metabolite quantification, there is no such digestion phase and so quantitative measurements are obtained.

Nevertheless, some care is required for absolute quantification. Indeed, in a cell extract, the enzymatic reactions can continue, leading to changes in concentrations. It is thus necessary to find ways to stop the reactions [8]. Generally, this is done by growing cells on a membrane filter, allowing nutrients to diffuse through the filter when it is disposed in an agarose plate. To quench the metabolism, the filter is removed from the plate and dropped in a dish containing cold organic solvent. This temperature drop stops the reactions so the concentrations of metabolites can remain fixed while the solvent denatures the enzymes. This conclusion however assumes that the metabolite of interest are stable; if a metabolite spontaneously becomes degraded, this effect has to be measured and used to correct the estimate of the quantification.

2.4.3 Measurements of fluxes

Metabolic fluxes are not measured as one would measure a flux going through a pipe, one needs to measure rates of chemical transformations... Ideally one would like to follow atoms and how their belonging to a metabolite type changes. A flux is the agregation of all the chemical changes from one given substrate to one product. At present it is not possible to track rates in individual cells, instead one follows what happens in a population of cells. To follow the belonging of atoms to metabolites, it is necessary to label them. In practice this relies on using isotopic markers, incorporating these into specific metabolites [23].

In one of the main approaches for flux measurements, one grows cells in a medium having a mix of different isotopomers (short for isotope isomer). When the growth has reached the steady state, the fluxes and the intra-cellular concentrations do not change in time. (Extra-cellular changes can be avoided too by working in a chemostat rather than in batch.) At this point, a sample is extracted to analyse the metabolic content of the cells. Typically, one measures the distribution of isotopomers among the amino acids because they are in large amount. Different biosynthetic pathways convert metabolic precursors into amino acids, each pathway potentially using different ways to re-organise atoms. When chosen appropriately, the isotopic labeling of the resulting amino acids contain much information about the fluxes through the different biosynthetic pathways of interest. For instance, the labelling pattern of the different amino acids can tell one about the relative contributions! of the associated fluxes while the growth rate of the culture can help provide absolute quantification of fluxes.

The most widely used label or "tag" is carbon ^{13}C, a stable isotope of carbon. ^{12}C is the most common isotope, while ^{13}C accounts for 1% of the natural carbon. ^{13}C contains seven neutrons which makes it a bit heavier than ^{12}C, containing six neutrons, and has the same charge. The two types of carbon atoms behave (almost) identically in chemical reactions but isotopomers are discernible with mass spectrometry. In most studies [23, 34], it is common to use different isotopomers of the substrate on which the bacteria are grown because not all the carbon atoms are passed to the same molecules. Glucose is often used under the forms: [U-^{13}C]glucose (uniformly labelled), [1-^{13}C] glucose (labeled only on the first carbon), and unlabelled.

To describe the flux inference from the steady-state isotopic content of the amino acids, I will use a toy example that mimics the two EMP and ED pathways of glycolysis presented in Fig. 2.6 B. The carbon source used to feed this pathway is composed of 20% [1-^{13}C]glucose and 80% unlabelled glucose. We

use the isotopomers of the valine amino acid as a reporter for the concentration of pyruvate; the valine variants have different masses which can be detected in a mass spectrometer, allowing one to measure their relative abondances as presented in Fig. 2.6 A. The choice of [1-^{13}C] labeling instead of [U-^{13}C] has a purpose: for such a labelling, the two pathways produce differently labeled pyruvate. The reaction v_2 produces pyruvate labelled on the third carbon whereas v_5 produces pyruvate labelled on the first carbon. With a [U-^{13}C] labeling, the two pathways would have instead produced the same la! belling of pyruvate. Along with the unlabelled pyruvate, the isotopomers contribute to six forms of valine. The proportion P_k^{Val} of the k^{iest} form of valine is obtained by summing the probability that two forms of pyruvate are involved in the reactions of formation:

$$P_k^{Val} = \sum_{\substack{i,j \\ \mathrm{Pyr}_i+\mathrm{Pyr}_j \to \mathrm{Val}_k}} P_i^{Pyr} P_j^{Pyr}$$

The toy system gives then an analytically solvable set of equations but in real analyses the data may be noisy and so one must deal with that. The solution is to find the set of pyruvate ratios that best fit the P^{Val} vector using a least square fit for example. It goes without saying that the sum of the proportions for any given metabolite is one.

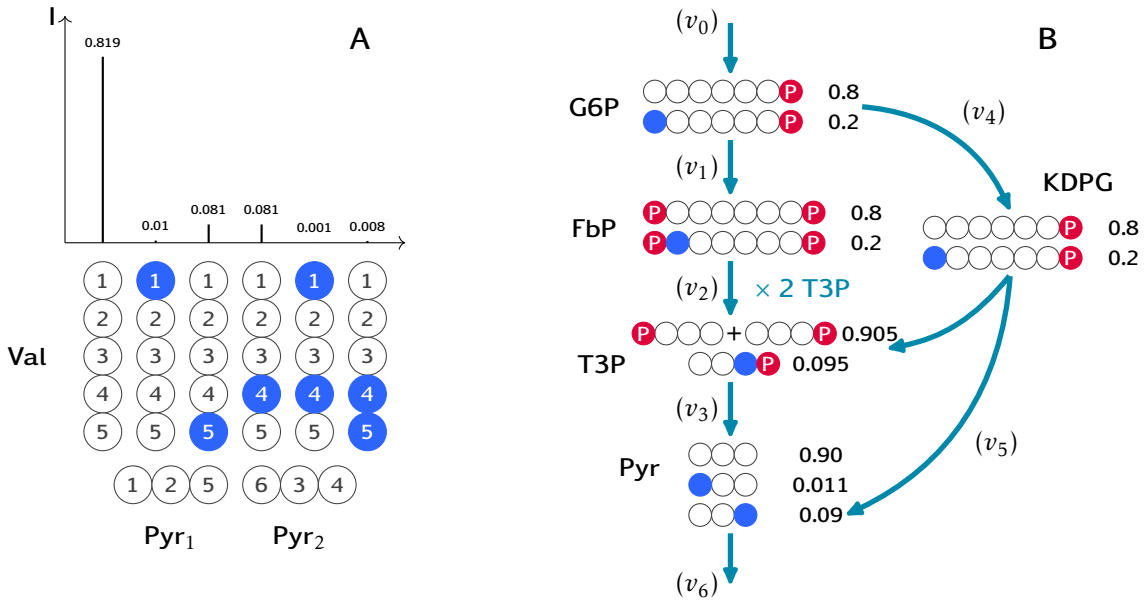


Figure 2.6: Flux inference for a simplified glycolysis. **A** The Valine (Val) amino acid consumes two pyruvate precursors to be produced; Val carbon content is mapped on the precursor pyruvates. The mass spectrometry signal intensity ratio is displayed for each of the possible isotopomers produced given the use of a [1-^{13}C]glucose marker. **B** Artificial network in which pyruvate is produced via two alternative pathways. The ratio of valine isotopomers follows from the metabolic proportion displayed. Labelled carbons are colored in blue, and phosphate groups are coloured in red to help carbon visualization. The steady-state fluxes compatible with such proportions are $v_1 = 0.9\,v_0$ and $v_4 = 0.1\,v_0$.

Based on the metabolite isotopomer distribution, it is generally possible (provided an appropriate labeling strategy) to calculate ratios of fluxes, i.e.,

$$P_k^M = \frac{Flux_{\to M_k}}{Flux_{\to M}}$$

where the numerator is the flux associated with the labeling k and the denominator is the total flux (all labellings included). Consider for example the case of the production of T3P in Fig. 2.6 B. One obtains

the explicit formula:

$$P_{3C}^{T3P} = \frac{v_2 \times P_{1C}^{FbP}}{2 \times v_2 + v_5}$$

Similarly, one may search for steady-state fluxes so that the data of isotopomer proportions is best reproduced.

The method described above allows one to find *ratios* between the different fluxes. To get absolute flux values, the nutrient uptake or growth rate are needed. This can be done in several ways; for instance, one can measure the glucose uptake using kits, or one can measure the rate of growth by measuring the time dependence of the cellular dried weight or of the optical density in the medium.

3

Characteristic times in metabolic networks

3.1 Introduction

Networks have been used to model systems involving large numbers of components, agents, or species [1]. Of particular interest are the effects arising in such systems either because of out-of-equilibrium dynamics or through equilibrium phase transitions. Collective effects are generally associated with slow dynamics, *i.e.*, characteristic times that are much larger than the microscopic times associated with elementary processes. In the present work our focus is on the emergence of large characteristic times in *reaction* networks close to their steady state. There are many ways to define a characteristic time in a dynamical system. The simplest is via the asymptotic relaxation towards the steady state [5, 40], relaxation which often will be exponential. If so, the amplitude of the perturbation or "distance" to the steady state will decay as $\exp(-t/\tau)$ at very long times, from which one then defines τ to be the *relaxation time*. Although in familiar situations τ is the longest characteristic time, our goal here is to investigate cases where much larger times can arise. Our study focuses on reaction networks for specificity, but our framework is more generally applicable to any system.

Reaction networks involve species that can transform one into another. If the species are molecular, one can get insights into the dynamics of the system by introducing an isotopic *tracer* and by following in time its incorporation into the different molecular species [53]. Assume that the reaction network is in contact with outside reservoirs, and let t_t be the time the tracer takes to exit the system. Surprisingly, the mean of t_t, corresponding to the lifetime in the system [32, 78] of the tracer (and sometimes called the mean residence time of the tracer), can be much *greater* than τ. The object of our work is to understand such a possibility, pointing in particular to the danger of assuming that τ is the main and longest characteristic time in these systems. For pedagogical reasons, we will begin by treating one-dimensional networks because an in-depth analytical treatment is feasible there, from which one can easily understand the influence of network size. We will then study more general systems using reaction networks published by other authors. In all cases, we compare the behaviors of *four* characteristic times in these systems, investigating the causes that can render them non informative or make their ratios diverge.

3.2 Models and Methods

3.2.1 Networks, molecular species and associated reactions

A metabolic network consists of a set of reactions and associated metabolites. It is convenient to represent such a network as a graph where the nodes are associated with metabolites; these are linked together by edges when there is a reaction that includes them as substrate and product. Such edges may be uni or bi-directional, accounting for the reversibility of the associated reaction. Let there be N metabolites M_i ($i = 1, ..., N$) and define C_i as the concentration of M_i. We are interested in the dynamics of the C_i, *i.e.*, how these quantities change with time and in the corresponding fluxes through the different reactions. Specifically, we shall study the dynamics close to the system's steady state and we shall probe the associated characteristic times. To facilitate the mathematical understanding of these times, we shall first focus on a particular kind of network consisting of a linear chain of reactions. In that situation, we order the metabolites from 0 to $N+1$ where the metabolite M_i is the product of reaction R_i whose substrate is metabolite M_{i-1}:

$$M_0 \overset{v_1}{\leftrightarrow} M_1 \overset{v_2}{\leftrightarrow} ... \overset{v_N}{\leftrightarrow} M_N \overset{v_{N+1}}{\leftrightarrow} M_{N+1} \tag{3.1}$$

The metabolites M_0 and M_{N+1} reside in infinite reservoirs at the two extremities of the chain so their concentrations are constant. By convention, the forward direction in such a chain goes from M_0 to M_{N+1}. Once understood the characteristic times in this system, we shall use the insight thereby gained to probe the situation in more realistic metabolic networks having branches and loops.

Reactions transform metabolites into other metabolites but it is necessary still to specify the actual kinetics. When a reaction happens spontaneously, without the need for a catalyst, it can be modelled by a mass action rate law (MA) where the flux is given by

$$v_i^{MA} = a_i C_{i-1} - b_i C_i \,. \tag{3.2}$$

To be specific, one can consider using the standard convention whereby concentrations are measured in Moles per litre and fluxes in Moles per liter per second. The parameter a_i (resp. b_i) is then the probability per second that a molecule of metabolite M_{i-1} (resp. M_i) spontaneously transforms into a molecule of metabolite M_i (resp. M_{i-1}). Note that Eq. 3.2 gives the total flux which is the forward flux minus the backward flux.

In practice, one is often interested in catalysed reactions where the spontaneous rates are terribly low. For instance, in biochemistry, most reactions are catalysed by enzymes; the catalysis allows for rates that can be enhanced by a factor of 10^{10} or more. For any such enzymatic reaction, the rate may be limited by the amount of enzyme and is no longer entirely proportional to metabolite concentration. Generally, the relation between substrate concentration and reaction rate grows linearly at low concentrations and then saturates at high concentrations of substrate. The reaction kinetics in this situation are typically modelled by the so-called reversible Michaelis-Menten-Henri (MMH) law [31]. In the case of a reaction involving one substrate and one product, the flux is given by

$$v_i^{MMH} = \frac{\alpha_i \frac{C_{i-1}}{K_i^{(S)}} - \beta_i \frac{C_i}{K_i^{(P)}}}{1 + \frac{C_{i-1}}{K_i^{(S)}} + \frac{C_i}{K_i^{(P)}}}. \tag{3.3}$$

Here, α_i is the maximum rate in the forward direction, reached when the substrate is in large excess and the product is absent. Similarly, β_i is the maximum rate in the backward direction. The maximum forward rate is proportional to the enzyme concentration and is often decomposed as $\alpha = k_{cat}E$ with E being the enzyme concentration and k_{cat} the maximum number of reactions catalysed by one molecule of enzyme per unit of time. $K_i^{(S)}$ and $K_i^{(P)}$, called the Michaelis constants respectively for substrate and product, are characteristic concentrations which set the scale for when the reaction becomes saturated in substrate or in product. For a MMH reaction in the absence of the product, $K^{(S)}$ is the concentration for which the rate is at half of its maximum value.

3.2.2 Determining steady states

When a physical system is not driven by outside forces, it goes to its equilibrium state where all net reaction fluxes are 0. In the context of our one dimensional model, that can only arise if the free energies of the two reservoirs are equal, corresponding to tuning the concentrations so that their ratio is the equilibrium one. Outside of that special case, the system will be out of equilibrium and concentrations will change in time until a steady state is reached which necessarily will have non zero fluxes. This steady state is generally unique if there are no regulatory processes but for our study to be completely general, we will not assume uniqueness of the steady state, we shall simply consider a stable steady state and investigate its characteristic times.

We have followed two approaches for determining steady states (leading to identical results):

1. solve the steady state equations $dC_i/dt = 0$ which we do numerically using root finding (routine "find-root" in Python). For any given boundary conditions, *i.e.*, concentrations C_0 and C_{N+1}, this leads to a list of steady-state concentrations $\overrightarrow{C^{ss}}$. It is necessary to check that the resulting steady state is linearly stable. This check can be performed using the linearised equations about the steady state. If $\overrightarrow{\delta C}$ is the (infinitesimal) difference between the actual concentrations and those in the steady state, one has

$$\frac{d\,\overrightarrow{\delta C}}{dt} = \mathbf{J}^{(\mathbf{c})}\,\overrightarrow{\delta C} \tag{3.4}$$

$$\mathbf{J}^{(\mathbf{c})}{}_{ij} = \begin{cases} A_i & \text{if} \quad j = i-1 \\ -(A_{i+1} + B_i) & \text{if} \quad j = i \\ B_{i+1} & \text{if} \quad j = i+1 \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

where the A_i and B_i are related to the terms entering Eq. 3.2 for mass action and Eq. 3.3 for Michaelis-Menten-Henri as specified in Table 3.1. $\mathbf{J^{(c)}}$ is the $N \times N$ Jacobian matrix with indices i and j going from 1 to N; the superscript c refers to the fact that it describes the (linearised) dynamics of (perturbed) *concentrations*. The steady state is stable if all the eigenvalues of the Jacobian have negative real part.

2. follow the concentrations using the dynamical equations (the system of ordinary differential equations specified by the kinetic laws) and extract the long time limit of the concentrations. This requires extrapolation, but generically takes one to a stable steady state.

3.2.3 Defining four characteristic times

- The first characteristic time is the *relaxation* time defined as $-1/\lambda_1^{(c)}$ where $\lambda_1^{(c)}$ is the real part of the leading eigenvalue of $\mathbf{J^{(c)}}$ having the largest real part. Because this time is defined via the linearised dynamics for the *concentrations* about the steady state, we shall refer to it as τ_c.

- The second characteristic time is the previously mentioned tracer *lifetime* (or mean residence time), which we denote by T_t. The motivation for introducing this quantity comes from tracer experiments in chemical networks where isotopic labels are used to follow atoms as reactions progress. Instead of introducing a perturbation to concentrations, this approach labels atoms of one metabolite M_k at $t = 0$ without changing any concentrations. In practice this labelling affects only a fraction of the molecules. The effect of this labelling is to leave the fluxes unperturbed as well. The system stays in its steady state, it is just that some of these concentrations become labelled. Note that when one labelled metabolite is tranformed into another, the labelling follows because the labelled atoms.

  Let us study the time evolution of the concentrations of these tracers $\overrightarrow{C_t} = \{C_{t,1}, C_{t,2}, \ldots, C_{t,N}\}$ (the subscript t is for *tracer*). As previously introduced, let $\overrightarrow{C^{ss}} = \{C_1^{ss}, C_2^{ss}, \ldots, C_N^{ss}\}$ be the steady state concentrations. Consider the reaction R_i and let ${\phi^{(f)}}_i$ be its forward flux and ${\phi^{(b)}}_i$ its backward flux in the steady state. Then the labelled concentration $C_{t,i}$ will include an incoming term given by the rescaled forward flux ${\phi^{(f)}}_i C_{t,i-1}/C_{i-1}^{ss}$ because all metabolite molecules (labelled or not) have an equal probability of participating in the reaction R_i. As a result, the dynamics of the tracer concentrations is

$$\frac{d\overrightarrow{C_t}}{dt} = \mathbf{J^{(t)}}\overrightarrow{C_t} \tag{3.6}$$

$$\mathbf{J^{(t)}}_{ij} = \begin{cases} \phi_i^{(f)}/C_j^{ss} & \text{if} \quad j = i-1 \\ -(\phi_i^{(f)}/C_{i-1}^{ss} + \phi_{i-1}^{(b)}/C_{i-1}^{ss}) & \text{if} \quad j = i \\ \phi_i^{(b)}/C_i^{ss} & \text{if} \quad j = i+1 \\ 0 \quad \text{otherwise} \end{cases} \tag{3.7}$$

  Note that these linear dynamics are exact even if $C_{t,i}$ is not infinitesimal. In general, the matrix $\mathbf{J}^{(t)}$ has no reason to be identical to $\mathbf{J^{(c)}}$. By exponentiating, one has the expression for the labelled concentrations at all times: $\overrightarrow{C_t}(t) = \exp(t\mathbf{J^{(t)}})\overrightarrow{C_t}(0)$. The lifetimeof the tracer is then taken to be the average over time of the probability of still being present in the system. This quantity depends the site at which the tracer is initially introduced. We define the tracer lifetime T_t as the largest such time when considering all possible initial sites:

$$T_t = \max_{\substack{\text{choice of initial} \\ \text{perturbation site}}} \left(\frac{\int_0^\infty |\overrightarrow{C_t}(t)| dt}{|\overrightarrow{C_t}(0)|} \right) \tag{3.8}$$

In this equation, $|\overrightarrow{C_t}(t)|$ is the norm of the vector. For our study, we use the L_1 norm ($|\overrightarrow{C_t}(t)| = \sum_i |C_t^i(t)|$) because it makes more sense for an atomic tracer which is conserved. Note also that T_t in Eq. 3.8 is the direct analog of the mean lifetime of a decaying positive *scalar* quantity; the norm allows one to extend the notion to a vector in a straightforward manner.

- The previous definition of lifetime of a tracer can be generalised to the lifetime of any quantity and in particular to a perturbation to steady-state concentrations. Suppose one introduces at $t = 0$ an infinitesimal perturbation in the concentrations, $\overrightarrow{\delta C}(0)$. Then according to Eq. 3.5, $\overrightarrow{\delta C}(t) = \exp(t\mathbf{J^{(c)}})\overrightarrow{\delta C}(0)$. In direct analogy with Eq. 3.8, the concentration lifetime T_c as

$$T_c = \max_{\substack{\text{choice of initial} \\ \text{perturbation site}}} \left(\frac{\int_0^\infty |\overrightarrow{\delta C}(t)| dt}{|\overrightarrow{\delta C}(0)|} \right) \tag{3.9}$$

  providing a third characteristic time of our system, referred to as the lifetime of a concentration perturbation. To be completely general, both here and for the tracer lifetimes, the vectors of concentrations should be taken as the deviations of their values from their long time limit. Indeed, if there were no reservoir and thus no exit possible of the atoms, the long time limit of the perturbation or tracer concentration would not be 0.

- Our fourth and last characteristic time is τ_t, defined as $-1/\lambda_1^{(t)}$ where $\lambda_1^{(t)}$ is here the real part of the leading eigenvalue of $\mathbf{J^{(t)}}$. It corresponds thus to the usual relaxation time but for the tracer molecules rather than for the metabolite concentrations, thus the subscript t.

3.3 Behavior of characteristic times in the one-dimensional network

As can be seen from the four characteristic times defined in the previous section, we distinguish two properties of a metabolic system: (i) the dynamics of an infinitesimal perturbation in the concentration of metabolites and (ii) the spreading and drift of tracers. Each of these properties can be considered when reaction kinetics are given by MA or MMH rate laws. In each case one can define both the standard relaxation time based on the asymptotic decay rate and a lifetime which measures the characteristic time needed for the system to return close to its steady state. In the case of a chain of reactions with the same kinetic parameters, the homogeneity allows us to obtain results analytically. For instance in the case of MA, the linearised dynamics ($\mathbf{J^{(c)}}$ and $\mathbf{J^{(t)}}$) are independent of the steady state chosen (that is the concentrations of M_0 and M_{N+1} do not enter) and the matrices are sufficiently simple for one to obtain in closed form the eigenvectors and eigenvalues. In the case of a MMH framework, when one performs the linearisation about the steady state, the resulting system is homogeneous only if the steady state itself is homogeneous, which requires that all the metabolites have the same concentrations. When this is the case, the steady state is again obtained in closed form. Furthermore, the eigenvectors and eigenvalues can be derived analytically, which gives us then the formulas for τ_c and τ_t. Unfortunately the study of the lifetimes T_c and T_t requires resorting to numerical methods to exploit Eqs. 3.9, 3.8. Nevertheless these algorithms are relatively straightforward as they reduce to calculating exponentials of the matrices $\mathbf{J^{(c)}}$ and $\mathbf{J^{(t)}}$ and performing the integrations in Eq. 3.9 and 3.8. For the initial perturbation, for simplicity we take $\overrightarrow{\delta C}(0)$ and $\overrightarrow{C_t}(0)$ to vanish everywhere except on the site at the center of the chain where the value is set to 1. For an even number of sites, there is no such centre so we average over the two most central sites.

3.3.1 Long transient times drive the gap between lifetimes and relaxation times

The integral in Eq. 3.8 depends on $\overrightarrow{C_t}(t) = \exp(t\mathbf{J^{(t)}})\overrightarrow{C_t}(0)$ which can be written using spectral decomposition as a sum of N terms, each term being associated with one eigenmode and having the time dependence $\exp(t\lambda_i^{(t)})$ where $\lambda_i^{(t)}$ is the associated eigenvalue. When $N = 1$, $\overrightarrow{C_t}(t)$ is a constant times a

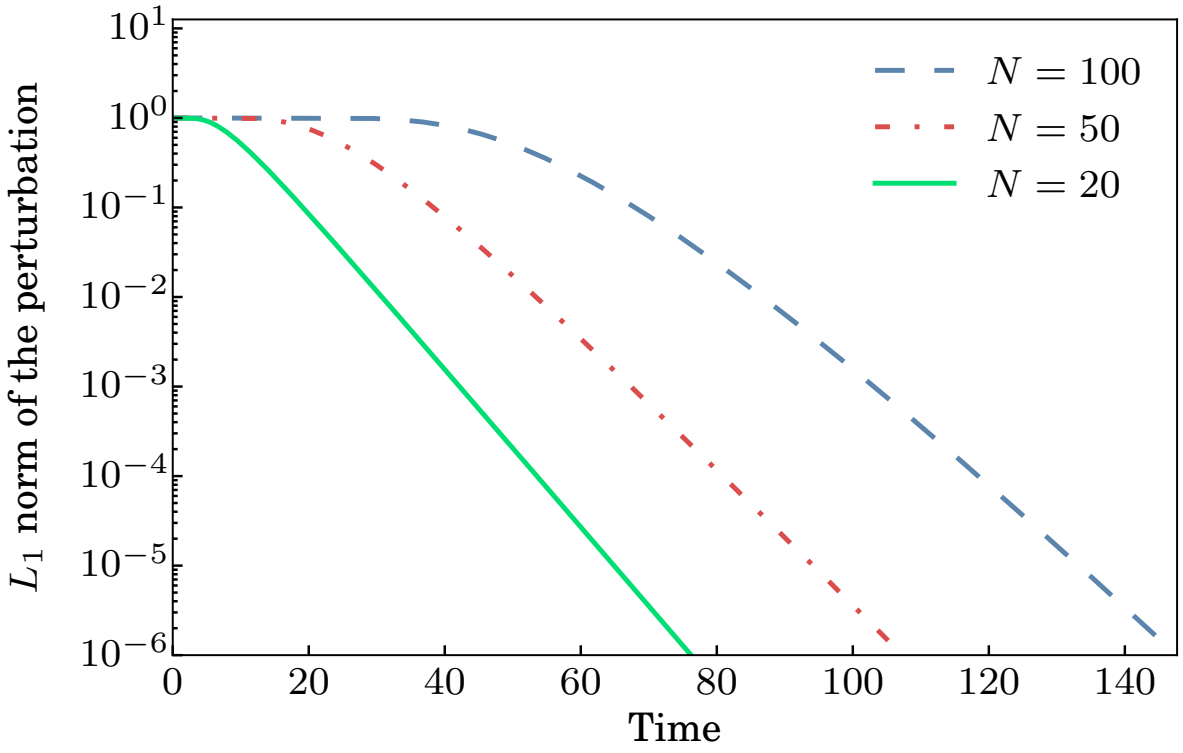


Figure 3.1: Decrease with time of $|\overrightarrow{C_t}|$, the L_1 norm of the vector of concentrations of a tracer. Identical results apply to $|\overrightarrow{\delta C}|$, the L_1 norm of the vector of perturbed concentrations. The initial perturbatation at $t = 0$ is localised at a site in the middle of the chain of reactions. The y axis is on a log scale so that one can see the asymptotic exponential decay as a straight line of slope $-1/\tau$: $\tau_{20} = 4.92$, $\tau_{50} = 5.65$, and $\tau_{100} = 5.78$. All N mass action reactions have $a = 2$ and $b = 1$. Shown are cases with $N = 20, 50$ and $N = 100$.

single decaying exponential. Plugging into Eq. 3.8 then reveals that $T_t = \tau_t$. The paradox whereby T_t can be much larger than τ_t arises only when $N \gg 1$. It is true that each of the N terms contributing to the spectral decomposition of $\overrightarrow{C_t}(t)$ decays in magnitude at least as fast as $\exp(-t/\tau)$ but that does *not* mean that the sum of these terms has that behavior on time scales comparable to τ. Indeed, the terms are not all of the same sign, and their cancellations can lead to long transients before the asymptotic behaviour (the exponential decay) prevails. To illustrate this, we show in Fig. 3.1 the L_1 norm of $\overrightarrow{C_t}(t)$ as a function of t in our toy model consisting of a chain with a's and b's identical across MA reactions. At large times, one sees the exponential decay (a straight line on this semi-log plot) but this asymptotic behavior may set in at times only much longer than τ itself. The cancellation at short times just mentioned is particularly striking: the curve is very flat for a very long time before it begins to decrease. That waiting time contributes to the large difference between T_t and τ_t and is associated with the transient time one must wait for tracer molecules to exit the system. Note that the property of having a very flat curve at initial times is due to the conservation of particles within the system, justifying our use of the L_1 norm instead of the L_2 norm.

3.3.2 Dependence of the characteristic times on N

Assuming the reactions to all have the same parameters and that the steady state is also homogeneous (cf. previous remarks), the relaxation time (be it τ_c or τ_t) can be obtained by using the translation invariance of $\mathbf{J}^{(\mathbf{c})}$ and $\mathbf{J}^{(\mathbf{t})}$. Each eigenvector is a product of a sine and an exponential. The formula for the eigenvalues leads to

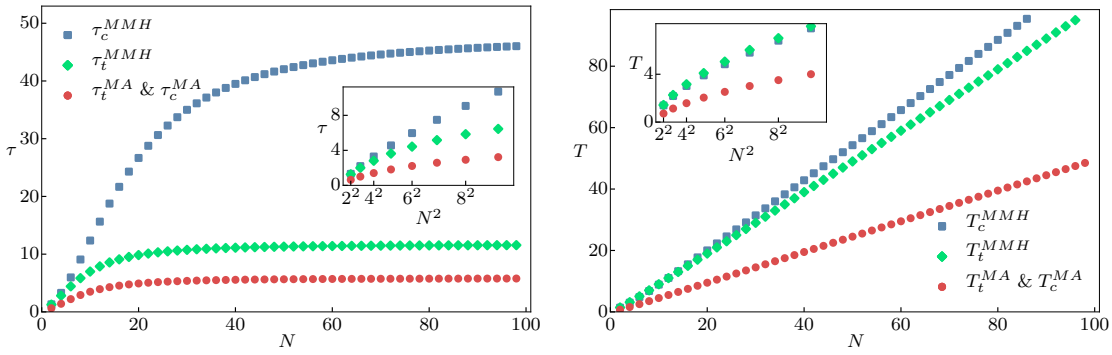


Figure 3.2: Relaxation times (*left*) and lifetimes (*right*) for chains between 2 and 100 metabolites long for a perturbation of concentrations and a tracer using the mass action or the Michaelis-Menten-Henri framework. Parameters: $a = 2$ and $b = 1$, $K^{(S)} = K^{(P)} = 2$ and $\alpha = aK^{(S)}$ and $\beta = bK^{(P)}$ so that the three conditions are comparable. The large N relaxation times are respectively $\tau_{c,lim}^{MA} = \tau_{t,lim}^{MA} = 5.83$, $\tau_{t,lim}^{MMH} = 11.66$, $\tau_{c,lim}^{MMH} = 47.66$. The transition sizes between a quadratic and constant or linear behaviour are $N_{c,cross}^{MA} = N_{t,cross}^{MA} = 7$, $N_{t,cross}^{MMH} = 7$, $N_{c,cross}^{MMH} = 17$. The insets illustrate the quadratic dependence on N for $N \ll N^{cross}$.

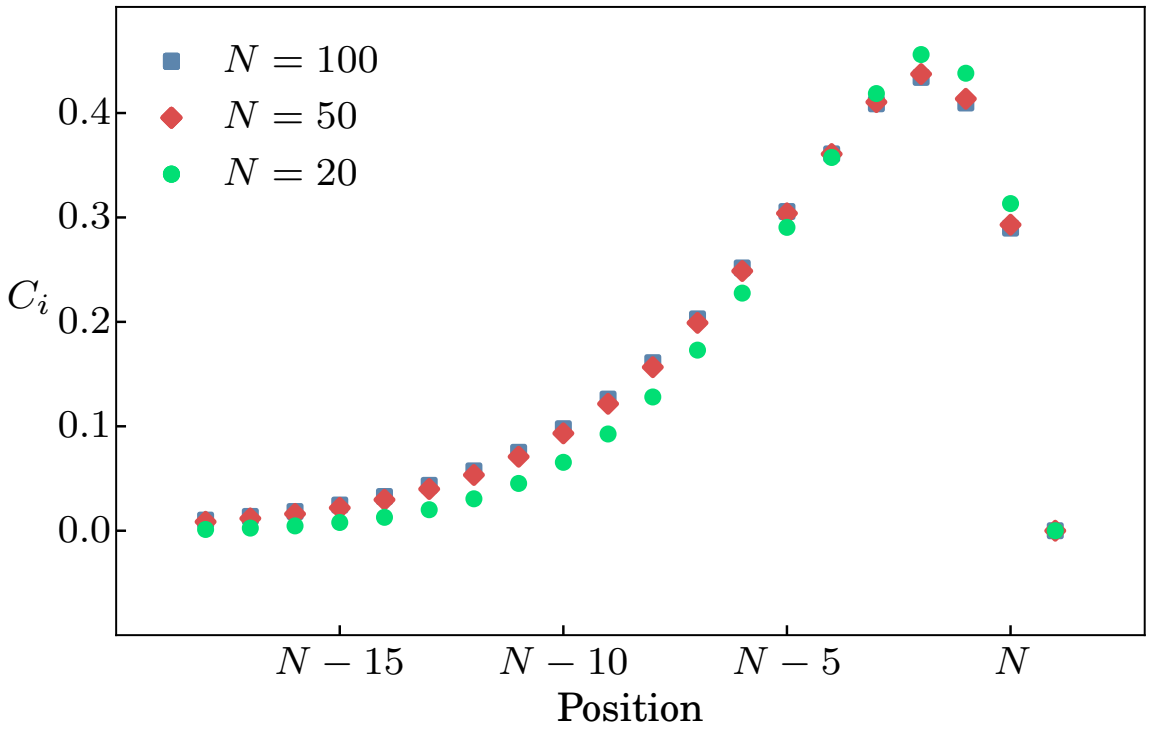


Figure 3.3: Leading eigenmode profile for the 20 last metabolites of the chain. Mass action parameters: $a = 2$, $b = 1$, and $N = 20, 50, 100$.

$$\tau = \frac{1}{A + B - 2\sqrt{AB}\cos\left(\frac{\pi}{N+1}\right)} \tag{3.10}$$

where the quantities A and B are the forward and backward probability of transition per unit of time

in the equations linearised about the steady state, entering in $\mathbf{J^{(c)}}$ for τ_c and in $\mathbf{J^{(t)}}$ for τ_t. They depend on whether one considers MA or MMH reaction kinetics and whether one considers a concentration perturbation or a tracer, the different cases being enumerated in Table 3.1.

Parameter	$MA-c$	$MA-t$	$MMH-c$	$MMH-t$
A	a	a	$\frac{\alpha-F}{K^{(S)}S}$	$\frac{\alpha}{K^{(S)}S}$
B	b	b	$\frac{\beta+F}{K^{(P)}S}$	$\frac{\beta}{K^{(P)}S}$

Table 3.1: Value of the A and B parameters for the four situations considered. F and $S = (1 + c^{ss}/K^{(S)} + c^{ss}/K^{(P)})$ are respectively the flux and the saturation factor at steady state in the network for the reactions, the system being by hypothesis homogeneous. The "c" (respectively the "t") appended to MA and MMH denotes that it is the perturbed concentrations (respectively the tracer concentrations) that are concerned.

The τs in the four cases are given by a standardised formula (Eq. 3.10), it is just that the proper A and B coefficients must be used. Note that for MA kinetics, $\mathbf{J^{(c)}} = \mathbf{J^{(t)}}$ so $\tau_c = \tau_t$. Furthermore, in both MA and MMH frameworks, when the relative difference between A and B is small, the τs exhibit two different regimes, one for small chains and one for long chains. For a short chain, $N \ll N^{cross} = \frac{2B\pi}{A-B}$, the characteristic times *grow quadratically* with the number of metabolites in the chain, a feature characteristic of diffusing systems for the simple reason that if $A = B$, the dynamics is purely diffusive. When N is much above this crossover value, τ_c and τ_t become independent of the chain length as can be seen directly by setting to 1 the cosine in Eq. 3.10. These two regimes are illustrated on the *left* of Fig. 3.2.

Note that the crossover size N^{cross} diverges as the inverse of $A - B$. Furthermore, in the context of MMH reaction kinetics, this crossover occurs for larger chain lengths when considering the dynamics of a concentration perturbation than when considering tracers because the *saturation* has the effect of reducing the difference between A and B. To illustrate these effects, we display on the *left* of Fig. 3.2 the relaxation times as a function of the chain length N for particular values of the kinetic parameters. As for MA, τ_c and τ_t do not increase asymptotically with N, the characteristic times become independent of the system size. To understand how this occurs, let us examine the leading eigenvector. Its entries depend exponentially on the index i of the node and so its profile is biased towards the largest indices. If the eigenvector with the largest eigenvalue becomes dominant, the major part of the deviation from the steady state is located on a few metabolites (about N^{cross}) at the end of the network. As illustrated in Fig. 3.3, if one increases the number of metabolites, that eigenmode just gets shifted to stay at the same position when measured from the end of the chain. As a consequence, increasing N does not affect the corresponding eigenvalue which determines τ. Thus τ_c and τ_t become independent of N at large N.

For the T_c and T_t lifetimes, we did not derive a closed form expression but one can still distinguish between two regimes. If $A - B$ is small, the behaviour for small N is diffusion-like so T_c and T_t increases quadratically with N. In contrast, for long chains, if $A \neq B$, one has a regime where T_c and T_t grow linearly with N. Similar arguments as for the relaxation times τ can be invoked to explain these two regimes. In small networks, the diffusion to the two sides of the chain dominates over the mean drift toward one end of the chain. In large networks, assuming $A > B$, most of the transient time dominating T_c and T_t is dedicated to the transport of the molecules to the $N + 1$ end, therefore that transient time is roughly equal to N divided by the drift velocity (which is proportional to $(A - B)$). We illustrate these different behaviours on the *rigth* of Fig. 3.2 where one sees again that the various cases behave similarly with the network length. (We already noted that for MA kinetics, $\mathbf{J^{(c)}} = \mathbf{J^{(t)}}$; as a consequence one has $T_c = T_t$ there, just as one has $\tau_c = \tau_t$.)

3.3.3 Effect of the saturation on the characteristic times

The major differences between MA and MMH come from the effect of the saturation. In the case of the MA rate laws, there is no saturation while saturation effects can be important in MMH kinetics. This difference can lead to much larger characteristic time scales in MMH than in MA whenever the concentrations are larger than $K^{(S)}$ or $K^{(P)}$. Furthermore, for highly saturated enzymes, the characteristic times

can be very different depending on whether one observes a tracer or a perturbation of concentration. Consider a reaction that is near saturation. Introducing a perturbation in the substrate will not change much the flux of that reaction and as a result it will take a long time to dissipate the perturbation away. On the other hand a tracer is essentially unaffected by saturation effects. Indeed, it is not because the reaction is saturated that the tracers cannot participate in the reactions. In effect, the tracers freely pass reactions that are saturated. The main consequence of this phenomenon is that in MMH τ_c can be much larger than τ_t (and T_c can be much larger than T_t).

To investigate quantitatively this phenomenon of particular relevance when interpreting kinetic properties from tracer measurements, let us increase saturation effects by reducing $K^{(S)}$. $K^{(P)}$ could also have been reduced, but when doing so, the flux in the network may reverse which unnecessarily complicates the analysis. Using the parameters of Table 3.1 in Eq. 3.10 for small values of $K^{(S)}$ gives the following analytical values for the dominant terms at small $K^{(S)}$ of the two relaxation times associated with a tracer (τ_t) and with a concentration perturbation (τ_c):

$$\tau_t \approx \frac{1}{\alpha} \tag{3.11a}$$

$$\tau_c \approx \frac{\left(\alpha+2\frac{\alpha+\beta}{K^{(P)}}-2\sqrt{\frac{\alpha}{c}\frac{\alpha+\beta}{K^{(P)}}\left(1+\frac{\alpha+\beta}{K^{(P)}}\right)}\cos\left(\frac{\kappa\pi}{N+1}\right)\right)^{-1}}{K^{(S)}} \tag{3.11b}$$

We see from these equations that τ_t becomes independent of the saturation while τ_c behaves linearly with $1/K^{(S)}$. Note that the saturation $S = 1 + c^{ss}/K^{(S)} + c^{ss}/K^{(P)}$ scales in the same way for small $K^{(S)}$. In Fig. 3.4 we show the dependence of the τs and the Ts on the saturation S for both a tracer and a concentration perturbation, assuming MMH rate laws. Not surprisingly, T_c is strongly affected by S, just as τ_c is.

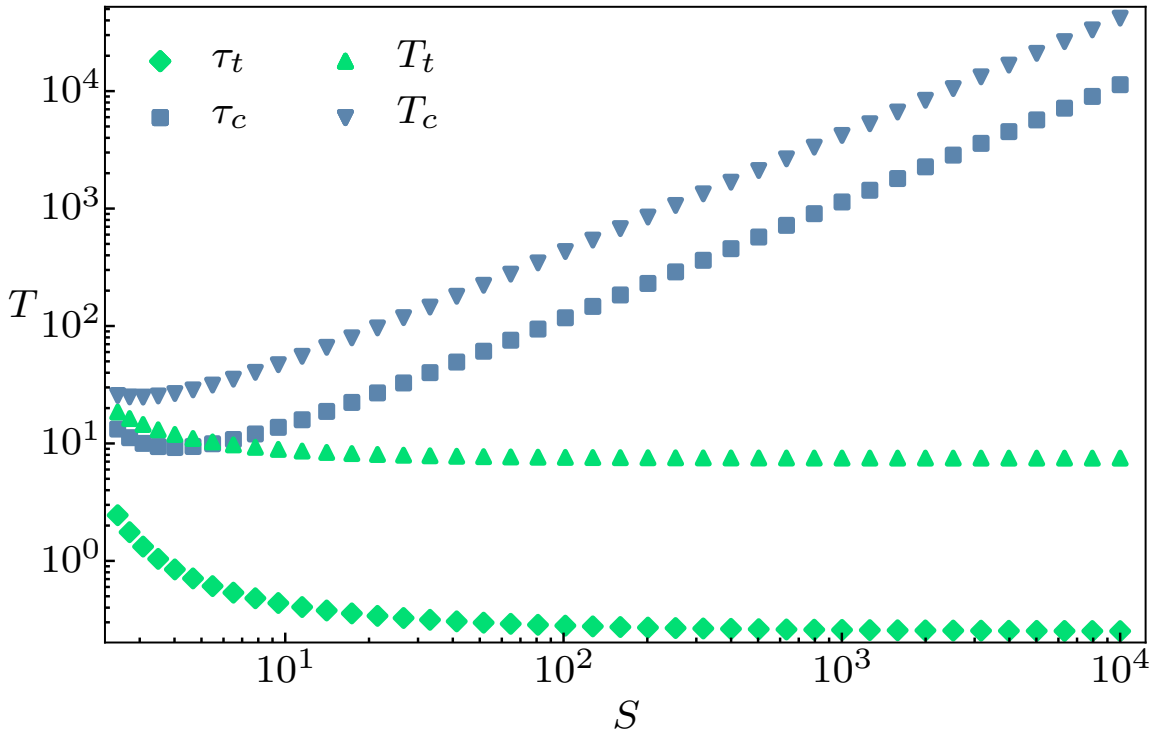


Figure 3.4: Relaxation times and lifetimes as a function of the saturation. Parameters: $\alpha = 4$, $\beta = 2$, $K^{(P)} = 2$, $N = 30$. To vary the saturations, the parameter $K^{(S)}$ is changed over a range going from 1 to 10^{-4}.

3.4 Behaviour of characteristic times in more general metabolic networks

3.4.1 Effects of disorder in the one dimensional chain

In the disordered (*i.e.*, heterogeneous) case we now consider, the rates "a" and "b" for the different reactions are taken to be independent random variables. Because every rate is a positive variable, we draw it from a lognormal distribution, *i.e.*, the natural logarithm of a rate r_i is distributed according to a Gaussian of mean μ and standard deviation σ. Consequently, the mean of r_i is $\bar{\mu} = \exp(\mu + \sigma^2/2)$ and its variance is $\bar{\sigma}^2 = (\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)$. We impose $\bar{\mu}$ to be equal to the value of the rate in the homogeneous case. An appealing feature of that way of introducing disorder is that the mean drift velocity of a marked molecule in Mass Action remains unchanged, being equal to its disorder average, $\langle a_i - b_i \rangle$. We are then left with the parameter $\bar{\sigma}$ which can go from 0 to ∞ and quantifies the intensity of the disorder. In practice, we use the same coefficient of variation for the "on" and the "off" reaction rates, corresponding to a single measure of intensity of disorder: $CV = \bar{\sigma}_a/a = \bar{\sigma}_b/b$.

For weak disorder, one expects little change in the values of the characteristic times $(\tau_c, \tau_t, T_c, T_t)$ compared to the homogeneous case. However, as disorder (CV) increases, the characteristic times typically do increase significantly. To identify the typical behaviour, we have determined these characteristic times for 10,000 realisations of the disorder and calculated the median times. We illustrate the associated results on the *left* of Fig. 3.5 for τ_c and τ_t in the case of Mass Action where those two quantities are equal. Increase is relatively mild (cf. the scales) at low CV but is more marked when CV is larger than 30%.

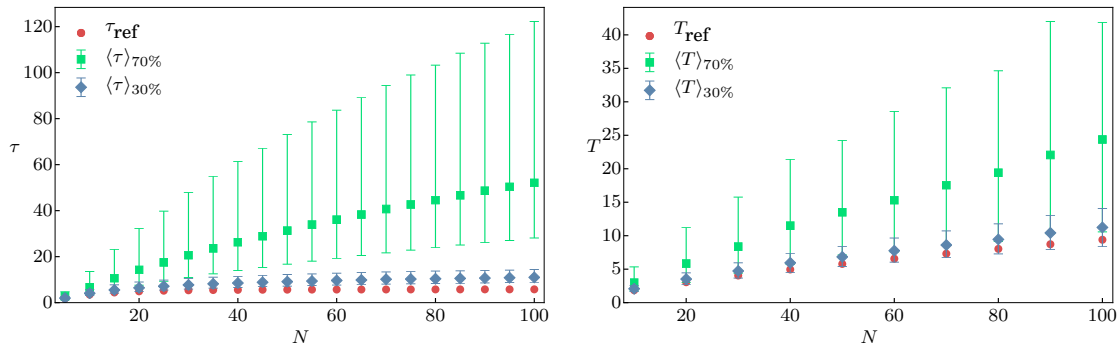


Figure 3.5: Median relaxation (*left*) and transit (*right*) times as a function of N for several intensities of disorder in the reaction rates as measured by their coefficient of variation CV for mass action kinetics. The error bars show the 68.2% confidence interval, value obtained by taking one standard deviation on both sides of the median of a Gaussian distribution. Parameters: $a = 2$, $b = 1$, $CV = 30\%$ and 70%.

Consider now the effects of disorder on the lifetimes. In Mass Action, $T_c = T_t$, even in the presence of disorder. We display on the *right* of Fig. 3.5 the dependence of these quantities on N for several values of CV and see that disorder has little effect as long as CV is small. This can be justified by noticing that the drift velocity of a molecule at site i is $a_i - b_{i-1}$ and its ensemble average (as in an annealed approximation) is the same as without disorder, namely $a - b$. At large disorder this argument fails because the quenched and annealed averages are very different. An extreme case can be seen from the fact that a large value of "a" at one site cannot compensate a small value at another site. At large CV, one sees significant effects of disorder. The reason should be clear: T_c and T_t are sensitive to unfavorable reactions (for instance where a is small) throughout the whole chain of reactions.

3.4.2 Networks with branches and loops

Although quite a few biosynthetic pathways include successive steps forming a chain of enzymatic reactions, the one dimensional systems considered so far remain toy models because in all known organisms, large scale biochemical metabolic networks have numerous branches and loops. It is thus necessary to consider how characteristic time scales might be affected by such structures. Rather than produce artificial networks including those features, it is more relevant to study directly the various kinetic models of metabolism that have been proposed in the literature. The repository "Biomodels" [36, 50] provides the gold standards for such models both because the models must past tests to be deposited and because their availability ensures that they can be compared to state of the art. Focusing further on those models that have been manually curated, we are left with only a handful of cases. The reason is that measuring kinetic constants of enzymes is a very difficult task so almost always when building a kinetic model the modeller has to use indirect methods to overcome the problem of dealing with many unknown parameters. We studied four of these models, published respectively in [14, 18, 56, 72].

For each of those four kinetic models, we first downloaded its SBML specification [36] from the repository and exported the ordinary differential equations into Python code that can be processed. Once in our format, we determined the steady state of the network of reactions and we then computed the matrices $\mathbf{J^{(c)}}$ and $\mathbf{J^{(t)}}$. The associated leading eigenvectors and eigenvalues were obtained using the inverse power method, thereby providing the values of τ_c and τ_t. Furthermore numerical integration was used to compute T_c and T_t according to Eqs. 3.8 and 3.9. The initial perturbation was taken to be localised at the first metabolite produced from the compound entering the network from the outside reservoir.

In Table 3.2 we provide the values of the four characteristic times for each of the Biomodels studied. The first model [72] contains the reactions for glycolysis in *S. cerevisiae* (baker's yeast). It has 17 reactions, mostly of the reversible MMH type, and there are 14 internal metabolites. Glucose is an external metabolite which enters the metabolism and then gets transformed. A total of 3 compounds can be excreted, all irreversibly. The characteristic times of this model are modest, from a few seconds to a few minutes. Further inspection shows that the ordering of these four values follows the same pattern as in our one dimensional toy model, namely

$$\tau_t < \tau_c < T_t < T_c \tag{3.12}$$

This can be justified as follows. First, $\tau_t < \tau_c$ and $T_t < T_c$ because a labelled atom is not subject to Michaelis-Menten saturation effects. The saturation of flux in a reaction may prevent a concentration fluctuation from being evacuated but it will not prevent labelled atoms from going through (participating to the flux). Furthermore, in our toy model, the τs are relatively insensitive to processes inside the network, they depend mainly on reactions close to the excreted metabolites, while the Ts depend on drift throughout the whole network and thus should be larger than the τs.

The other models follow quite closely but not exactly this same pattern (cf. Table 3.2). Model 2 contains the reactions for the glycolysis and the pentose phosphate pathway in *E. coli* [14]. It has 48 reactions and 17 internal metabolites, but we needed to remove the model's explicit time dependence to allow a steady state. The main difference with the model 1 is the organism considered and the glucose steady state uptake rate (3.1 $\mu mol.s^{-1}.L^{-1}$ compared to 1.5 $mmol.s^{-1}.L^{-1}$) but Eq. 3.12 is respected. Model 3 contains the glycolysis and the pentose phosphate pathway, but for a human cancer cell. It has 29 reactions and 34 internal metabolites. The glucose uptake, expressed per gram of cell dry weight ($0.17 mmol.s^{-1}.gcdw^{-1}$), cannot be compared to the two previous uptakes but most of the inequalities of Eq. 3.12 are satisfied.

Model 4 contains the reactions for the biosynthesis of purines in *E. coli* [18]. It has a total of 29 reactions and 18 internal metabolites. The main difference compared to the other three models is that the formalism uses kinetics that are neither MA nor MMH: the forward and backward rates of the reactions are fractional powers of the concentrations of the metabolites. Such fractional powers are often used phenomenologically to parametrise allosteric or regulatory effects; they have the drawback that the flux may rise very steeply when starting with low concentrations; although this may be the case for some regulatory processes, it can lead to a situation where a concentration perturbation will be evacuated more efficiently than a labelled atom. Such a possibility seems to be realised in this model

since in Table 3.2 one sees that $\tau_c < \tau_t$ and $T_c < T_t$. This model may have further pathologies as might be indicated by the huge values of all the four characteristic times.

time (s)	τ_c	τ_t	T_c	T_t
Model 1 [72]	15.	3.75	339	84.4
Model 2 [14]	120	95.2	2834	2210
Model 3 [56]	4.94	0.16	107	3.53
Model 4 [18]	$4.34\ 10^5$	$1.11\ 10^6$	$9.35\ 10^6$	$2.36\ 10^7$

Table 3.2: Value of the characteristic times τ_c, τ_t, T_c and T_t in seconds for the four manually curated models [14, 18, 56, 72] we have studied and that are available on the Biomodels repository [50].

4

Kinetic modeling of *E. coli* 's central carbon metabolism: an automatized construction methodology

Kinetic models of metabolism have typically focused on specific pathways or other small scale networks for which it has been possible to get a lot of quantitative information. In the absence of that kind of detailed knowledge, metabolic modeling has generally restricted itself to steady states where the balance of matter (implemented in the so-called Flux-Based-Analysis or "FBA") can provide insightful constraints, even though kinetic aspects are set aside. The main achievement of my thesis work has been to develop computational tools to construct – in an automatic fashion – kinetic models for large networks, even if there is not much information available for individual reactions. The present chapter is dedicated to the description of the workflow to go from the topology of the reactions all the way to the construction of a complete kinetic model.

4.1 The core network and its coupling to biomass production

The RESET project brings together a consortium of laboratories to study ways to reorient *E. coli* 's resources away from growth processes and towards production of metabolites of interest. For instance, metabolism from growth pathways, such as those producing amino acids (AA), should be (temporarily) shut off, and the precursors of those pathways should be channeled into other pathways synthesising metabolites that have economic value and that can be extracted. This project has been funded by the French PIA investment program over a period of four years. As a proof of concept, the first phase of RESET targets the metabolite glycerol. My contribution to that objective was to first develop a kinetic model of *E. coli* 's central metabolism and then to use it to test *in silico* whether reduction of flux towards the production of AA can lead to enhanced production and yield for the production of glycerol.

The main part of these three years of thesis work consisted in developing a satisfactory kinetic model of *E. coli* 's metabolism.

To begin that endeavor, it was necessary to decide on the appropriate level of detail to use for defining the model, and even more importantly, what set of reactions should be included. Clearly, tackling a genome-scale framework with over 700 reactions was neither necessary nor feasible: having so many reactions would lead to several times that many unknown parameters and consequently far too much uncertainty in the behavior of the model's characteristics. A more realistic approach would consist in focusing on the catabolic pathways (for breaking down the nutrients provided in the medium) along with the key biosynthesis pathways producing the main building blocks of the cell (AA, nucleic acids). However that appraoch would still involve a very large number of unknown parameters. Furthermore, those biosynthesis pathways incorporate regulatory processes that are often unknown and thus their modeling would necessarily involve a lot of adhoc choices. Fortunately, the "operating principles" of these pathways can be roughly summarized by the fact that they allow for flux when the final product (AA or nucleic acid in our case) is limiting. Such (simplified) regulatory logic can be implemented by exploiting the so called essential precursor metabolites which feed into these biosynthesis pathways; it is then possible to model this feeding-in by effective reactions going directly from these essential precursors to the biomass building blocks. In view of these constraints and possibilities, my choice settled on working with the central carbon metabolism along with just a few extra reactions required for the RESET project. This focus allowed me to develop and optimize automatic procedures for model building. But given this proof of concept, my approach should be useful for studying other metabolic systems and creating kinetic models even when experimental data is sparse.

4.1.1 The heart of the model: the CCM

Reducing the model to still fewer reactions could provide a toy system for developing our methodologies and I did investigate that approach to some extent, focusing for instance just on glycolysis. But for the purposes of the RESET project, in particular for understanding the possibilities of reorienting fluxes when demand for AA or nucleic acids drops, one cannot significantly reduce the complexity of the network: essentially all of the central carbon metabolism (CCM) must be included along with effective reactions describing biomass production. Since carbon is the central actor of the CCM, it is important that its stoechiometry be exact throughout the whole network. (Ideally all the reactions should be stoechiometrically exact, but sometimes reactions are not known with certainty, especially in genome-wide

models; the point here is that errors in reactions may have minor consequences, but in our system, any error in the carbon y would have disastrous consequences.) Even though this choice of focusing on CCM and a few extra reactions may seem logical and simple, it hides the fact that some metabolites involved in the reactions of the CCM are not produced there nor in our effective biosynthesis reactions. This situation is particularly frequent for co-factors, an example being NADH. This difficulty is ubiquitous, arising in all metabolic modeling work, it is always a challenge to go from a full metabolic model to a reduced one because of such "hanging" metabolites and even reactions. The cure is generally to consider that the concentrations of these particular metabolites is kept fixed, a procedure I will follow. Such concentrations may be either set (for instance from experimental measurements) or considered as parameters to be adjusted along with the many other ones associated with reaction rates.

Among all the parameters determining reaction rates, the equilibrium constants are probably the most studied ones. In fact, one has quite reliable estimates of their values, and so to use this information I decided to model all the CCM reactions as reversible. Such a choice makes the network less modular (recall that having irreversible reactions reduces feedback) and thus a bit more complex but it is also more realistic. In my modeling, only the fluxes exiting from the CCM are taken to be irreversible. These irreversible reactions include those going from essential precursors to secondary synthesis pathways, those associated with excreted metabolites such as acetate, and those involved in the production of glycerol (cf. the objectives of the RESET project). Glycerol is formed from DHAP via two enzymatic steps:

$$DHAP + NADH \rightleftharpoons Gly3P + NAD$$

$$Gly3P \rightarrow Gly + Pi$$

Since no other metabolite than Gly3P and GAP are produced from DHAP, the flux of glycerol is directly proportional to the concentration of Gly3P and thus is proportional to the flux exiting from DHAP. The complete set of reactions included in our model is presented in Tab.4.2.

For the kinetic laws, reaction dynamics were modeled using convenience kinetics for all the reversible reactions. For the irreversible reactions (associated for instance with effective reactions), there was a single substrate and so I used irreversible MMH rate equations. When a metabolite regulates a reaction, I model its action by multiplying the unregulated rate with a factor as given in Eq. 2.12.

4.1.2 Flux towards biomass

The set of fluxes going into biomass production are not directly incorporated into my model, instead I use as a proxy the fluxes originating from the biomass precursors. Since the model is fitted to physiological fluxes, I assume that the precursors feed into the biomass synthesis pathways "optimally". To be explicit, the precursors are taken to generate production of AA in the proportions given by the biomass composition formula. Furthermore, the produced AA feed into a pool used for protein synthesis. In effect, the process of translation (polymerization of AA to synthesize proteins) draws from this AA pool with the constraint that the fluxes from the different AA again obey the fixed proportions in the biomass formula. Suppose now one introduces a gedanken experiment where the rate of protein sythesis is reduced, e.g., by some manipulation of the cell's expression machinery as planned in the RESET project. The corresponding lowering of the rate of AA consumption would lead to an increase in AA concentration; the feedback inhibition of AA on the enzymes driving their own synthesis would then propagate this flux decrease all the way back to the CCM, inhibiting the reactions which consume the precursors feeding into biomass. This overall process is then easy to implement in my reduced CCM model. A more extensive description of this gedanken experiment and its treatment in silico will be presented in chapter 6.

4.2 Data available for building a kinetic model

The main source of difficulty in building a kinetic model is the very large number of reaction rates to describe, with respect to both their form and their parameter values. Even if one knew the *type* of kinetic law (Michaelis Menten, convenience kinetics, feedback inhibition, ...) appropriate for every reaction,

Name		Reaction		Effectors
PTS	Glc + **PEP**	$\rightleftharpoons$	**G6P** + **Pyr**	
Pgi	**G6P**	$\rightleftharpoons$	**F6P**	**PEP**(-)
Pfk	**F6P** + ATP	$\rightleftharpoons$	ADP + **FbP**	
Aldo	**FbP**	$\rightleftharpoons$	**DHAP** +**GAP**	α**Kg**(+)
Tis	**DHAP**	$\rightleftharpoons$	**GAP**	
Gdh	**GAP** + Phosph + NAD	$\rightleftharpoons$	**BPG** + NADH	
Pgk	**BPG** + ADP	$\rightleftharpoons$	**PGA3** + ATP	
Pgm	**PGA3**	$\rightleftharpoons$	**PGA2**	
Eno	**PGA2**	$\rightleftharpoons$	**PEP**	
Pk	**PEP** + ADP	$\rightleftharpoons$	**Pyr** + ATP	**FbP**(+)
Zwf	**G6P** + NADP	$\rightleftharpoons$	**GL6P** + NADPH	
Pgl	**GL6P**	$\rightleftharpoons$	**PGn**	
Gnd	**PGn** + NADP	$\rightleftharpoons$	**Ru5P** + NADPH +CO2	
Rpi	**Ru5P**	$\rightleftharpoons$	**R5P**	
Rpe	**Ru5P**	$\rightleftharpoons$	**X5P**	
TktA	**X5P** + **E4P**	$\rightleftharpoons$	**GAP** + **F6P**	
TktB	**R5P**+ **X5P**	$\rightleftharpoons$	**GAP** +**S7P**	
Tal	**S7P** + **GAP**	$\rightleftharpoons$	**E4P** + **F6P**	
Pdh	**Pyr** + CoA + NAD	$\rightleftharpoons$	CO2 +**AcCoA** + NADH	
Acs	Phosph + ADP + **AcCoA**	$\rightleftharpoons$	ATP + CoA + **Ace**	
Ppc	**PEP** + CO2	$\rightleftharpoons$	Phosph + **OAA**	**AcCoA**(+), **FbP**(-), **Mal**(-)
Cs	**OAA** + **AcCoA**	$\rightleftharpoons$	CoA + **Cit**	
Acn	**Cit**	$\rightleftharpoons$	**Icit**	
Icdh	**Icit** + NADP	$\rightleftharpoons$	NADPH +CO2 + α**Kg**	
Kgdh	α**Kg** + CoA + NAD	$\rightleftharpoons$	CO2 + **SucCoA** + NADH	**OAA**(-)
Stk	**SucCoA** + Phosph + ADP	$\rightleftharpoons$	CoA + **Suc** + ATP	
Sdh	**Suc** + Ubi	$\rightleftharpoons$	**Fum** +UbiH2	
FumA	**Fum**	$\rightleftharpoons$	**Mal**	**Cit**(-)
Mdh	NAD + **Mal**	$\rightleftharpoons$	**OAA** + NADH	
Mae	**Mal** +NAD	$\rightleftharpoons$	CO2 + **Pyr** + NADH	**G6P**(+), **OAA**(-), **AcCoA**(-), **Fum**(-)
Edd	**PGn**	$\rightleftharpoons$	**KDPG**	
Eda	**KDPG**	$\rightleftharpoons$	**GAP** + **Pyr**	
out_{G6P}	**G6P**	$\rightarrow$	$\emptyset$	
out_{DHAP}	**DHAP**	$\rightarrow$	$\emptyset$	
out_{R5P}	**R5P**	$\rightarrow$	$\emptyset$	
out_{E4P}	**E4P**	$\rightarrow$	$\emptyset$	
out_{PGA3}	**PGA3**	$\rightarrow$	$\emptyset$	
out_{PEP}	**PEP**	$\rightarrow$	$\emptyset$	
out_{Pyr}	**Pyr**	$\rightarrow$	$\emptyset$	
out_{AcCoA}	**AcCoA**	$\rightarrow$	$\emptyset$	
out_{Ace}	**Ace**	$\rightarrow$	$\emptyset$	
out_{OAA}	**OAA**	$\rightarrow$	$\emptyset$	
$\text{out}_{\alpha Kg}$	α**Kg**	$\rightarrow$	$\emptyset$	
out_{SucCoA}	**SucCoA**	$\rightarrow$	$\emptyset$	

Table 4.2: Reactions included in the model, the reversible reactions are presented with bidirectional harpoons and irreversible reactions are symbolised by unidirectional arrows. The bold metabolites are the variables of the model and the other are used as the model parameters. The effectors are ppresented in the left column dans denoted with a (+) for the activators and a (-) for the inhibitors.

there are very few experimental measurements on the reactions allowing one to extract the associated kinetic parameter values. As we will see, a kinetic model of the central carbon metabolism in *E. coli* can easily involve more than 200 unknown parameters. To provide estimates of the parameter values in such a model, one must go beyond information produced by measurements on isolated reactions: the only practical approach today is to use information associated with the *systemic* properties of the whole set of reactions. Put simply, one must use quantitative knowledge at the level of the whole system to infer by *computation* the properties of its individual components. As a consequence, I will exploit as much as possible available measurements performed on the whole network, measurements which depend of course on all individual components but in a complex way. The framework proposed is quite general but in my thesis work I implement it for the central carbon metabolism of *E. coli* .

4.2.1 Equilibrium constants (k_{eq})

As already mentioned in chapter 2, the equilibrium constant of a reaction does not depend on its enzyme, it is a function only of the substrates and products. This means that the k_{eq} of a reaction is the same in all organisms provided the environmental conditions are the same (same pH, same ionic force). The value of k_{eq} can be calculated if the standard Gibbs potential energies of formation are known for each of the metabolites and products of the reaction. Sometimes not all required standard energies have been measured; in such cases, it may be possible to infer missing values by considering a path of reactions of known energy change involving one such metabolite. (Such an approach exploits experimental measurements of changes of Gibbs energies associated with reactions, changes that can be obtained via calorimetric measurements.) As a result, the equilibrium constants of most reactions are actually quite well determined, in particular for those of the central carbon metabolism.

Once the topology of the metabolic network is settled upon, *i.e.*, once one has selected the set of reactions (with their substrates and products) to consider, a literature search should provide a great many of the associated equilibrium constants (or equivalently the Gibbs energies of the reactions). There are necessarily some uncertainties in these values coming from experimental effects or lack of control of the environmental conditions (ionic forces etc.). Because measurements are redundant (for instance formation energies determine reaction energies but these are measured anyway), it is possible to improve the estimates by taking all the measurements (including the redundant ones) and fitting the formation energies of the different metabolites (which are non redundant) to these data. This procedure reduces then the uncertainties just as multiple measurements will reduce a statistical error. Furthermore, the approach also provides an estimate of the *uncertainty* in the estimated value of each formation energy. (The eQuilibrator website calculates these uncertainties by assuming a Gaussian distribution for the quantities to estimate.)

Another major advantage of working with formation energies rather than with the k_{eq} (or equivalently with the *reaction* energies) is that thermodynamic consistency is automatically built into the formalism. To illustrate this phenomenon, I show in Fig 4.1 the kind of inconsistency one will inevitably encounter if one works with the equilibrium constants (reaction energies) instead of the formation energies. Since databases provide many k_{eq}, one is tempted to use these values directly, but in fact it is much safer to reinterpret them in terms of formation energies to ensure thermodynamic consistency.

The data used for in our modeling for the formation energies come from the eQuilibrator platform [58]. This web platform uses the *component contribution* method, presented earlier, to combine experimental data and data calculated from group contribution to produce a consistent database for reaction and formation standard energies. It also allows the user to model the influence of the pH and of the ionic strength. I wrote a script to have the platform compute the formation energies of the metabolites included in the model. Although the pH and the ionic strength can vary depending on the growth conditions, the standard energies are computed for a constant pH of 7.5 [80] and a fixed ionic strength of 250 *mM* [66] which seem to be reasonable values for an average bacterium.

As noted when Tab.4.2 was introduced, even if some metabolite concentrations are taken as parameters in the model (they cannot vary in time), the script also searches for their corresponding formation energies since these values are needed to compute the reaction energies.

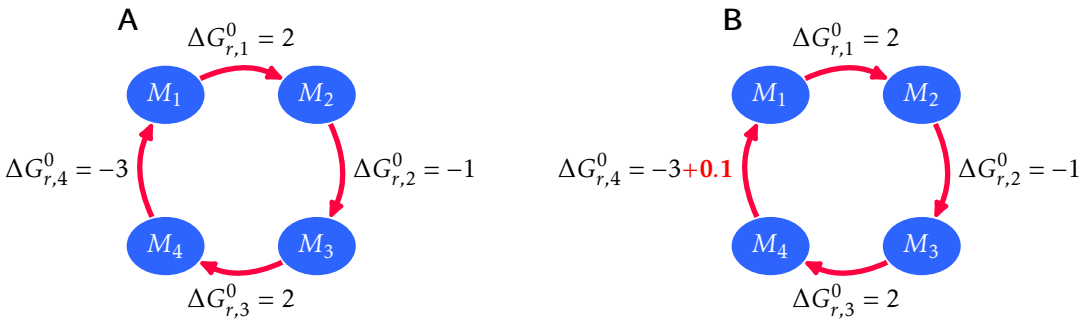


Figure 4.1: Standard reaction energies without (A) and with (B) a small error in the estimated values of some free energies. The values are taken for instance from a database where the second database has $\Delta G^0_{r,4} = 3.1$ instead of 3 (arbitrary units). In the case where the database has no statistical error or bias (A), the model is thermodynamically consistent and it is possible to close the loop: formation energies provide a true potential energy landscape. In the case B (the value given by the database has a small error), there is a spontaneous production of energy. A metabolite M_1 with an initial energy of 0 will end up with an energy of 0.1 after one cycle, so the system described by the second database is thermodynamically inconsistent.

4.2.2 Steady-state reaction fluxes in the CCM

As mentioned previously, although kinetic measurements on individual reactions are scarce, many systemic measurements have been performed in metabolic networks. Our first example of this is the flux through various reactions when the whole network is operating. A sensible objective is to ensure that the model built in this thesis reproduces *in silico* these systemic results. To be specific, I developed my model to reproduce as much as possible a set of steady-state fluxes (referred to as the target fluxes) provided by Haverkorn van Rijsewijk & al. [34]. That dataset was one of the first to provide high quality measurements of fluxes in the *E. coli* CCM. In that article, the authors measured the fluxes in the three pathways (glycolysis, PPP, and TCA) of the central carbon metabolism. To do so, they used ^{13}C labeling in the *E. coli* K-12 BW25113 bacterium grown on a minimal medium with glucose at 4 $g\,l^{-1}$. The results they obtained are presented in Fig. 4.2. The best estimate for each flux along with its 95% confidence interval are listed in B.

There are several ways to account for the uncertainties in the fluxes given the experimental means and confidence intervals. Such accounting will be necessary when fitting the model to the experimental data. One way is to assume fluxes have a Gaussian distribution. I chose instead to use a log normal distribution of the fluxes, and I will do the same for the concentrations (see later). Thus I take the natural logarithm of each flux, $\log(F)$, to follow a Gaussian distribution. Then the fitting of the model will be done assuming that F has a log-normal probability distribution:

$$P(F) \propto \frac{1}{F\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log(F)-\mu}{\sigma}\right)^2} \tag{4.1}$$

To have the same mean and standard deviation values as those provided by Haverkorn van Rijsewijk & al., I set the parameters μ and σ of the log-normal distributions as

$$\begin{aligned} \mu &= \log(E) - \frac{\sigma^2}{2} \\ \sigma^2 &= \log\left(1 + \frac{V}{E^2}\right) \end{aligned} \tag{4.2}$$

where E is the mean value of the distribution in [34] and V is the corresponding variance equal to s^2, s being taken so that $1.96s$ is half of the 95% confidence interval size.

The motivation for the choice of the log-normal distribution (compared for instance to a Gaussian distribution) may not seem obvious but it is important. Errors coming from measurements with physical

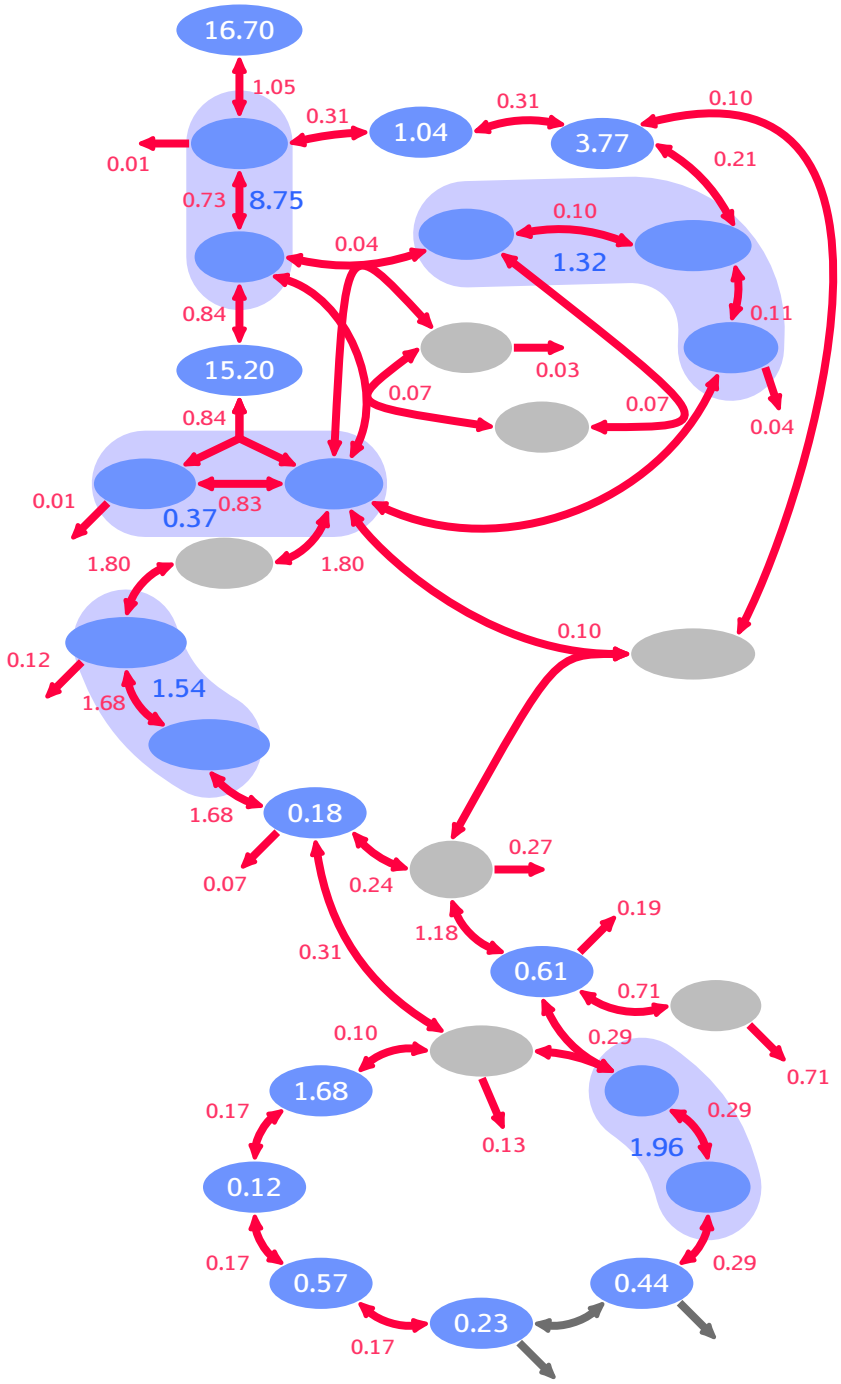


Figure 4.2: Steady-state flux values (Red) and concentrations of metabolites (Blue) from the literature [7, 34]. Some metabolites were undiscernable in the measurements and thus appear as clustered into a pooled measurement (surrounded by light blue). Grey symbols stand for quantities for which no measurements are available. The fluxes are expressed in $mM\,l^{-1}$ while concentrations are expressed in mM.

instruments are best thought of as being relative rather than absolute. The fundamental reason for this is that instruments are designed to be able to span over multiple orders of magnitude and thus must

intrinsically have larger absolute uncertainties when the value of the measurement grows. With our log-normal choice, it is the relative error that matters, and as a result there is a big difference between a flux which is 100 times too small and a flux which is just 10 times too small. Such a clear contrast would *not* arise when using a Gaussian distribution as can be seen directly from the Gaussian formula for the probability. Clearly, in the experimental measurements, such factors of 10 are important, and using the log-normal distribution is natural, translating simply the notion that the *relative* error is what matters. Without such a choice, the optimization procedure for adjusting the model parameters will not penalize null or near null fluxes, a feature which is unacceptable.

Note that there are cases of reactions where no steady-state flux measurements are provided; we then leave these fluxes as free, though their value will be constrained by the other fluxes because of the steady-state conditions.

4.2.3 Steady-state concentrations of metabolites in the CCM

Having a kinetic model reproducing target fluxes is nice but further contact with experimental results at the systemic level is possible. Specifically, the model will be further constrained if it must also reproduce experimental values of *metabolite concentrations*. To constrain the range of possible concentrations, I used Bennett & al.'s study [7] that quantified an important number of metabolites in the *E. coli* K-12 NCM3722 strain. That is not the same strain as used by Haverkorn van Rijsewijk & al. but it is close. Furthermore, both groups performed their experiments on a minimal medium implying the necessity for cells to produce all amino acids *de novo* which is appropriate for our modeling. Lastly, the glucose concentrations used by these two groups is similar, at about 3 $g\,l^{-1}$.

It is fair to criticize the path I followed here because the two data sets to be used, one for fluxes and the other for concentrations, were obtained in slightly different conditions and for non-identical strains. Nevertheless, one should keep in mind that the idea behind my proposed method is to use systemic information as a way to enforce constraints on the many parameters of the model and that the quality of the data is a strong issue only if the goal is to provide reliable predictions from the model. Our aims concerning predictions to extract from the model are at the qualitative level only and to some extent follows from the absence of high quality data. In that situation, the predictions may be unreliable but our work is still relevant for providing a methodology that can be useful and predictive given better data sets.

Bennett & al. measured many different types of metabolites, including metabolites arising in the CCM, cofactors and amino acids. Some of these metabolites are difficult to measure because they have a short lifetime (in particular, in vivo they are rapidly transformed by the enzymatic reactions). This difficulty can lead to biases in the estimated concentrations and so generally the concentrations of metabolites are not determined to high accuracy. An additional difficulty arises for certain metabolites which cannot be uniquely identified. For example, consider G6P and F6P. These two metabolites have the same mass and so cannot be distinguished by mass spectrometry which measures the ratio m/z. The same difficulty arises for other groups of isomers in the CCM. Instead of constraining these molecules individually, the measurements constrain them as a pool, namely the only accessible quantity is the sum of their concentrations. This type of missing information is illustrated in Fig. 4.2 where we display the values provided by Bennett & al.'s. In that figure, one sees these pooled metabolites where only the sum of concentrations is known.

In [7] the concentrations are reported along with a 95% confidence interval as presented in B. Similarly to what I did for the exploitation of fluxes, for fitting the model I will use a probability distribution (or likelihood) of values of a metabolite (or pool) concentration that is of the log-normal type (cf. Eq. 4.2). The parameters of the log-normal distribution (mean and variance) will be set in agreement with the Bennett results. For the metabolites whose concentrations have not been measured, it is possible to just let them be free as was done for fluxes. But in contrast to that case where the steady-state condition put strong constraints on remaining fluxes, here the concentrations remain quite uncertain. Thus we build on the intuition that outlier concentrations are unlikely to arise. To define what "outlier" is quantitatively, the simplest approach is to compare to the empirical distribution produced by the measurements of Bennett. I thus determined the distribution of concentrations taken from his measurements as shown in Fig. 4.3. This distribution tells us about the range allowed for a concentration if it is

to be considered typical of the CCM. I then fitted this distribution to a log-normal form, leading to the two associated parameters μ and σ. That distribution is then considered as a prior for the unmeasured metabolite concentrations.

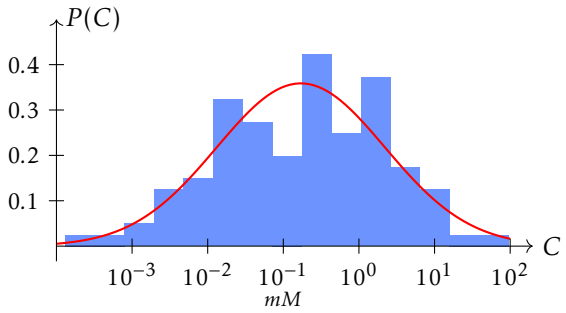


Figure 4.3: Distribution of measured metabolite concentrations in the CCM, from Bennett & al. [7] (in blue). The distribution of the logarithm of the concentration is approximated by a Gaussian (in red). The parameters of the Gaussian are $\mu = -1.768$ and $\sigma = 2.561$.

Looking at Fig. 4.3, the goodness of the fit is questionable. But the idea is not to describe the detailed features of the concentration distribution but rather to have a range within which concentrations of *E. coli* CCM metabolites are expected to lie.

4.2.4 Prior distributions for the kinetic parameters k_{cat} and K^M

It is often the case in kinetic modeling that when parameters are available in the literature, the modeler fixes them to the published values. This has the nice consequence of facilitating the fitting of the remaining parameters, in effect reducing the complexity of the model. However such an approach is appropriate only if the value of the published estimate is sufficiently accurate. In many cases, there can be uncertainties in such values by more than factors of two. Another drawback is that the value of the parameter may vary with experimental conditions; if such variations are significant, it becomes important to fit the model to homogeneous experimental situations [41]. A particular example of this follows from the fact that kinetic parameters are generally measured in vitro [28, 45], using purified enzymes and thus non-physiological conditions, and such an approach can produce severe discrepancies with *in vivo* measurements [28]. A confirmation of this difficulty arises when looking at the literature: a lot of variability occurs accross different publications for a single kinetic parameter value. For example, K^M_{FbP} for reaction Zwf is measured to be 1.76 mM in one publication and 0.17 mM in another. Similar trends are seen also for the published values of k_{cat}. The conclusion is that it is difficult to decide how much confidence to give to an experimental value if (1) no confidence interval is provided or (2) when few groups have reproduced the measurement.

In this thesis I adopted a method in which the k_{cat} and K^M are given prior probabilities. The prior distribution for K^M (resp. k_{cat}) is generated from the data concerning all strains of *E. coli* 's CCM present provided by the BRENDA [68] and EcoCyc [47] databases and from published papers [6, 7]. Inspired by the method used in the previous section for assigning prior distributions to unknown concentrations, I fitted the the logarithm of the parameter values (from these previous works) to Gaussians as illustrated in Fig. 4.5. I then used these log-normal distributions to assess the probability for a parameter to take on any given value.

This technique, inspired by Bayesian frameworks, leaves some flexibility for the parameter values but nevertheless guides the optimization algorithm towards realistic ranges of values. It also allows for an *a posteriori* examination of the fitted values because if some values do turn out to be outliers, one may identify them and search for metabolic or evolutionary arguments to justify such unexpected behavior.

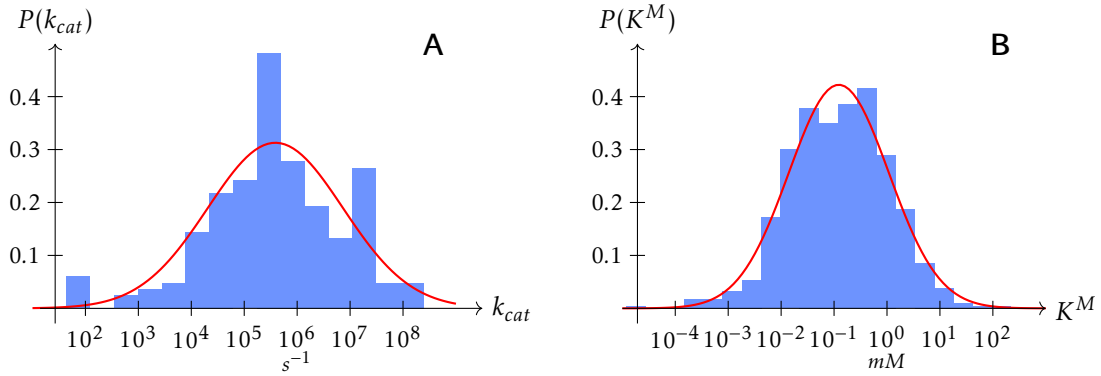


Figure 4.4: The empirical distributions for the parameters k_{cat}(A) and K^M(B) found in the literature for *E. coli* 's metabolism (in blue). In red, the fit to a log-normal distribution, used as the prior in the model optimization. The parameters of the log-normal distributions are $\mu_{k_{cat}} = 2.94$, $\sigma_{k_{cat}} = 12.9$, $\mu_{K^M} = 2.18$, and $\sigma_{K^M} = 2.10$.

4.2.5 Enzyme concentrations and V_m

As mentioned above, measurements of metabolite concentrations must overcome the short lifetimes of metabolites in the presence of enzymatic reactions. In contrast, measuring enzyme (protein) concentrations encounters no such problem because proteins are quite stable. Mass spectrometry is again the preferred technique, but quantification requires using labeled species to benchmark abundances and refined analyses of peptide spectra. In spite of these obstacles, enzyme quantification is commonly realized and a lot of data are available in the literature. Just as for fluxes and other metabolic quantities, the concentrations of enzymes may depend strongly on the physiological state of the cell. For example, the cell will not invest its resources (protein building blocks and energetic costs for amino acid polymerization) for the production of TCA enzymes, an aerobic pathway, if the conditions are anaerobic. This cautionary note is no different from what was given above for both fluxes and metabolite concentrations. In the presence of reliable data, clearly it is a good strategy to include the quantification of enzyme abundances into the model, but in the absence of measurements one can let the concentration be free or use a weak prior distribution when fitting the model.

The concentration of an enzyme affects the rate of a reaction only through a multiplicative factor; in effect, it just scales up the reaction rate associated with one molecule of enzyme. From the cell's viewpoint, for given conditions, the relevant quantity is neither the enzyme quantity E nor the catalysis rate k_{cat} but their product, equal to the maximum velocity $V_m = k_{cat} \times E$. Using explicitly these two quantities instead of their product in the optimization of the model leads to undetermination: a change in k_{cat} can be compensated by a change in E unless priors are used. Since we do have priors for k_{cat}, this undetermination is absent and it is not necessary to impose any prior on E. But since my model has so many parameters, any additional information or priors is of interest; thus I do introduce a prior distribution for E. To do so, I followed the same logic as when introducing priors for metabolite concentrations or for kinetic parameters. Given enzyme concentrations from different sources [74,75], I generated a reference distribution of abundances and fit it to a log-normal distribution. The corresponding parameters are $\mu = -4.95$ and $\sigma = 1.81$. Note that depending on the published work used, the data are given as a number of proteins per cell, or in dimensions of mass of protein per cell, or mass of protein per mass of cell dry weight ($gCDW$). All data were converted to mM units. When necessary, I used the enzyme masses from EcoCyc [47], and the cells were assumed to have a constant volume of 0.6 μm^3 [71] and mass of 0.28 pg [37].

Now a technical point must be explained which allows us to reduce the number of parameters to fit in the model. Begin by noticing that the kinetic properties of the model depend only the product of the enzyme distribution and of the catalysis constant, not on the separate values. V_m is equal to $k_{cat}E$, and so the idea is to fit directly V_m, not k_{cat} and E separately, thereby reducing the complexity of the

model. To perform this simplification, we need the prior of V_m in terms of the priors of k_{cat} and E. This is achieved starting with

$$\log(V_m) = \log(k_{cat}) + \log(E)$$

Because of the priors, $\log(k_{cat})$ and $\log(E)$ are two randomly distributed variables and we take them to be independent. The expression for the probability density of $\log(V_m)$ is then

$$P^V(\log V_m) = \int_{\log E_{min}}^{\log E_{max}} P^E(\log E) P^{k_{cat}}(\log V_m - \log E) d\log E$$

where P^V, $P^{k_{cat}}$, and P^E are respectively the distributions of $\log(V_m)$, $\log(k_{cat})$ and $\log(E)$. P^V is nothing else than the the convolution of $P^{k_{cat}}$ and P^E. Because of that, we choose P^E to also be Gaussian (just like $P^{k_{cat}}$) and as a result, P^V is also. Naturally, their means and variances satisfy

$$\begin{aligned} \mu_V &= \mu_{k_{cat}} + \mu_E \\ \sigma_V^2 &= \sigma_{k_{cat}}^2 + \sigma_E^2 \end{aligned} \tag{4.3}$$

Thus from characteristics of the distributions for $\log E$ and $\log k_{cat}$, it is easy to fully characterize the distribution of V_m. Then the model involves only the V_m parameters instead of both the k_{cat} and E parameters, and that simplification leads to higher computational efficiency.

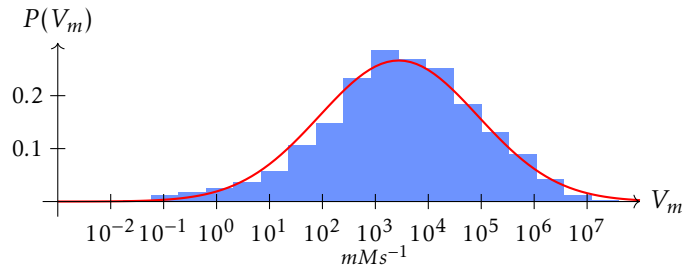


Figure 4.5: The distribution of the composite parameter V_m. The data, in blue, correspond to all the possible products of the experimental k_{cat}s with the experimental Es. The parameters of the prior used in the model, in red, are obtained from Eq. 4.3: $\mu = 3.45$ and $\sigma = 7.92$.

4.2.6 Mean passage time in the CCM

The steady-state fluxes and concentrations arising in the metabolic network are precious systemic pieces of information that our model must reproduce. However, if the application of the model involves only steady states, it might have been sufficient to use methods like FBA that require far less parameters have already proven their efficiency for describing large models. Hence to take full advantage of the kinetic nature of our model, it is of interest to reproduce time course data.

A standard approach in kinetic modeling is to fit time series data, for instance after a metabolic perturbation has been applied [14, 15, 43, 81]. A perturbation may correspond to applying a pulse of glucose in an otherwise steady state situation, or injecting a poison which halts one of the reactions. The goal behind most of these kinetic models is to understand the detailed regulation and physiology of the cell on short time scales following the perturbation. To obtain a satisfactory model, it is often necessary to fix the concentrations of certain components in a particular way, e.g., one might use a polynomial to parametrize the time course of certain metabolites, in particular cofactors such as ATP, ADP, NAD, NADH, etc because they are involved everywhere in the cell and the model cannot account for their production or consumption. These imposed time course for such components are not compatible with a model aiming at representing steady states.

In my case, the model is not built with the objective of giving a perfect agreement with the detailed variation of metabolic pools during the first phase of some perturbed dynamics. Instead, I want it to

describe reasonably standard steady-states and also steady states reached after imposing a time independent perturbation. For this purpose, I need time courses on longer time scales. Some such measurements were obtained from Meier & al. [53] who measured the quantity of ^{13}C labelled isotopomer for different metabolites of the CCM for more than one minute after a pulse of uniformly labeled glucose. I will thus exploit such (time-dependent) data to optimize my model. More generally, any data, be-they from time dependent behavior or steady states, can be used to further refine the model parameters.

Mean passage time

In chapter 3, I defined two important characteristic times for a tracer, the relaxation time τ and the mean transit time – also called the mean residence time – T. Let the linearized dynamics about a stable steady state be described by the Jacobien J. Then the standard definition of the system's relaxation time is minus the inverse of the eigenvalue of J that is closest to 0. Furthermore, the mean residence time is given by the following integral:

$$T = \int_0^\infty \frac{|\vec{C}(t)|}{|\vec{C_0}|} dt \tag{4.4}$$

In this equation $\vec{C}$ is the tracer concentration at time t ($\vec{C_0}$ is $\vec{C}$ at the initial time 0 when the labeling is applied). Estimating the mean transit time requires the full tracer concentration vector at each measurement whereas Meier measured only a few metabolites from the CCM (DHAP, Pyr, AcCoA, and Ace). Thus the mean transit time cannot be inferred and another characteristic time must be defined. To do so, I define the mean passage time (MPT) which is the average time at which labelled metabolites are found within a given metabolite:

$$MPT_i = \frac{1}{N_i^{tot}} \int_0^\infty N_i(t) dt \tag{4.5}$$

where N_i^{tot} and $N_i(t)$ are respectively the total quantity of labelled carbons passing within the state (metabolite) i and the quantity of carbons that are in the state i at time t. There still remains a problem with the normalisation of the MPT because the number of molecules arriving in state i is not necessarily the number of labelled carbon introduced initially, some may have exited from the network before and the quantity is defined in arbitrary units in [53]. To obtain a quantity that can be extrapolated to any system, I choose to define the transformed MPT^* which is the MPT scaled by a factor N^{tot}/N^{max}, N^{max} being the number of labeled molecule at the maximum of the presence curve. This new characteristic time can then be written as:

$$MPT_i^* = \frac{1}{N_i^{max}} \int_0^\infty N_i(t) dt \tag{4.6}$$

Note that MPT_i^* is proportional to MPT and does not depend on the unit in which the number of labeled particles is expressed. For instance, if instead of counting the number of carbon atoms within state i one looks at the concentration of metabolite i, one would obtain the same value for MPT^* by replacing the N_i and N_i^{max} by the concentrations C_i and C_i^{max} in Eq. 4.6 provided that the number of labeled carbons is shared equitably among the molecules i.

Exploitation of the time series

The experimental time series are a bit noisy and some concentration points are even indicated as negative so it is hard to exploit them by measuring the area under the experimental points. I prefered to approximate the MPT^* with the closed form functions $MPT^*(t) = at^{b-1}e^{-ct}$, inspired from a gamma distribution which is usually used for modeling these kinds of time phenomena. When using a linear model for the tracer dynamics, the modes decay according to an expoential law and the sum of exponential laws give a gamma function. The characteristic relaxation time for a given metabolite in this parameterization is $1/c$. The experimental data and inferred MPT^* are presented in Fig .4.6. The MPT^* is specific to each metabolite while the relaxation time is a global quantity. In a linear approximation of a model about the steady state, only the dominant τ is easy to observe. I chose the greatest measured value of τ as the dominant one, more specificaly $\tau_{DHAP} = 13.1$.

The fact that these data are quite noisy and that the gamma distribution does not always fit perfectly the time series (e.g., for pyruvate) pushes me to consider that these times are orders of magnitude rather

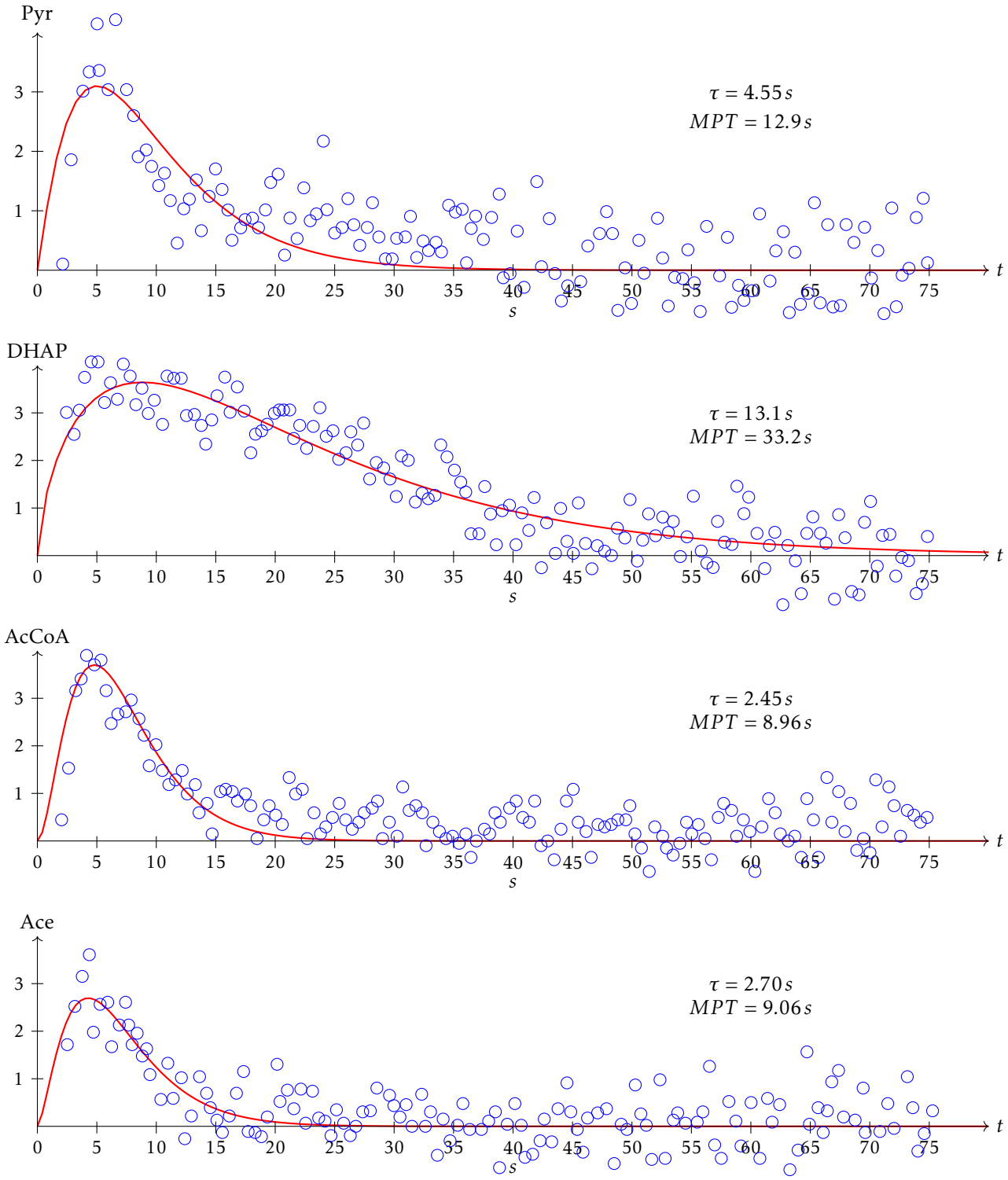


Figure 4.6: Fit of experimental time series [53] for the quantities of Pyr, DHAP, AcCoA and Ace via a function $MPT^*(t) = at^{b-1}e^{-ct}$. The quantities of labeled metabolites are in arbitrary units. The maximum relaxation time is found for DHAP: $\tau_{DHAP} = 13.1$.

than accurate estimations of characteristic times of the system. To account for this lack of accuracy, I evaluated the probability that a choice of parameters in my model provides agreement with the previous four mean passage times as well as with the relaxation time. I did so using five Gaussians centered on the measured values and of standard deviation inferred from the data.

4.3 Optimization of the model

The optimization of the model consists in finding the "optimal" set of parameters that (i) reproduces the experimental data, consisting of Haverkorn van Rijsewijk's fluxes and Bennett's concentrations, and (ii) is consistent with the prior distributions that has been explicited earlier in this chapter. Starting from an initial model, the procedure implements a search for improved parameter sets until no further improvement is found. For this approach, one needs (1) a quantitative measure of the goodness-of-fit, and (2) an algorithm for searching the parameter space and performing the associated optimimization.

4.3.1 A score for the goodness-of-fit

For a given set of values of the model's parameters, we need to quantify the deviations of the properties of the model from the experimental data and from the prior distribution. This is most easily implemented using a quasi-Bayesian framework where the goodness-of-fit can be associated with a score which plays the role of a likelihood. One has a prior distribution for every parameter and we already provided a probability that measures the accordance of the model's characteristics (fluxes and concentrations, but this can be extended also to characteristic times) with experimental data. Each of these distributions is built based on experimental results presented earlier in this chapter. It is then possible to define a proxy for the likelihood of the parameter values of our model. This likelihood is not a true likelihood, it is more like a score or penalty for deviations between the model's predictions and experimental results or the prior information. This (pseudo) likelihood has a contribution for each type of constraint imposed. For example, the term associated with the constraint on concentrations is given by

$$\mathscr{L}(C^m) = \mathscr{L}(\log C^m) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\log C^M - \mu)^2}{2\sigma^2}}$$

where C^M is the model's predicted value for the concentration, and the right hand side is associated with the log-normal ($LogN(\mu,\sigma)$) distribution of measuring experimentally this quantity.

In the CCM model, the total likelihood is obtained by multiplying these terms for all the measured quantities. For a given set of parameters $\vec{p}$, one thus has

$$\begin{aligned}\mathscr{L}(\vec{p}) = \quad & \frac{1}{K} \prod_i e^{-\frac{(\log F_i - \mu_i)^2}{2\sigma_i^2}} \prod_j e^{-\frac{(\log C_j - \mu_j)^2}{2\sigma_j^2}} \prod_k e^{-\frac{(\log V_{m,k} - \mu_k)^2}{2\sigma_k^2}} \prod_m e^{-\frac{(\log K_m^M - \mu_m)^2}{2\sigma_m^2}} \prod_n e^{-\frac{(\Delta G_{f,n} - \mu_n)^2}{2\sigma_n^2}} \\ \times \quad & e^{-\frac{(MPT^*_{Pyr} - \mu_{Pyr})^2}{2\sigma^2_{Pyr}}} e^{-\frac{(MPT^*_{DHAP} - \mu_{DHAP})^2}{2\sigma^2_{DHAP}}} e^{-\frac{(MPT^*_{AcCoA} - \mu_{AcCoA})^2}{2\sigma^2_{AcCoA}}} e^{-\frac{(MPT^*_{Ace} - \mu_{Ace})^2}{2\sigma^2_{Ace}}} e^{-\frac{(\tau - \mu_\tau)^2}{2\sigma_\tau^2}}\end{aligned} \tag{4.7}$$

See Tab. 4.3 for more information on the different variables arising in these expressions.

The objective of the optimisation algorithm is then to search for the set of parameters that results in the highest likelihood $\mathscr{L}(\vec{p})$. However, the logarithm being a monotonically increasing function, it is more convenient to work with the log-likelihood function since it involves a χ^2 of the logarithm of the measured quantities or of the quantities themselves depending on whether they are distributed

K	Normalization constant
F_i	Steady state flux of reaction i
C_j	Steady state concentration of metabolite number j
$V_{m,k}$	Value of the k^{iest} parameter V_m
K_m^M	Value of the m^{iest} parameter K^M
$\Delta G_{f,n}$	Formation energy of metabolite number n
MPT^*_{Pyr}	Mean first passing time of metabolite Pyr
MPT^*_{DHAP}	Mean first passing time of metabolite $DHAP$
MPT^*_{AcCoA}	Mean first passing time of metabolite $AcCoA$
MPT^*_{Ace}	Mean first passing time of metabolite Ace
τ	Dominant relaxation time for the exponential decay of an isotopic tracer
μ_x,σ_x	Parameters of the distribution for variable x

Table 4.3: Variables involved in Eq. 4.8.

log-normally or normally. With the same quantities as previously, this new score is:

$$\begin{aligned}\log\mathscr{L}(\vec{p}) = & -\sum_i \frac{(\log F_i-\mu_i)^2}{2\sigma_i^2} - \sum_j \frac{(\log C_j-\mu_j)^2}{2\sigma_j^2} - \sum_k -\frac{(\log V_{m,k}-\mu_k)^2}{2\sigma_k^2} - \sum_m \frac{(\log K_m^M-\mu_m)^2}{2\sigma_m^2} \\ & -\sum_n \frac{(\Delta G_{f,n}-\mu_n)^2}{2\sigma_n^2} - \frac{(MPT^*_{Pyr}-\mu_{Pyr})^2}{2\sigma^2_{Pyr}} - \frac{(MPT^*_{DHAP}-\mu_{DHAP})^2}{2\sigma^2_{DHAP}} - \frac{(MPT^*_{AcCoA}-\mu_{AcCoA})^2}{2\sigma^2_{AcCoA}} \\ & -\frac{(MPT^*_{Ace}-\mu_{Ace})^2}{2\sigma^2_{Ace}} - \frac{(\tau-\mu_\tau)^2}{2\sigma_\tau^2}\end{aligned} \tag{4.8}$$

The optimal model is the one having the minimal value for $\log\mathscr{L}$. Note that the normalisation constant has been removed since it is the same for every set of parameters.

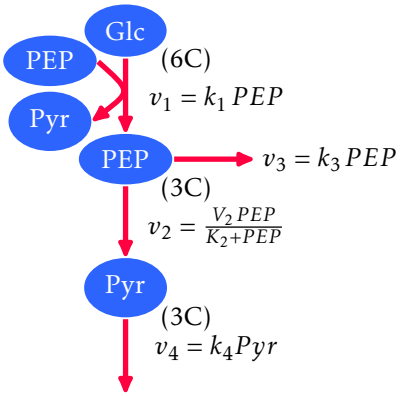
4.3.2 Initializing the model parameters before optimization

Before searching for a model that optimises the $\log\mathscr{L}$ function, it is necessary to first define an initial set of parameters so that the associated model has a sensible steady state. Unfortunately testing a random set of parameters is not a good strategy: in practice one generally finds that the steady state has all fluxes and concentrations at 0. Note that to sustain a steady state, fluxes need to produce as much of metabolite as they consume. A particularly instructive example is the coupling of PEP between the two reactions PTS and Eno. Eno must regenerate the pool of PEP at the same rate as PTS consumes it, otherwise the system simply depletes and converges to being empty. In other word, given the maximal stoechiometry 2PEP:1glucose, glycolysis is allowed to divert to other pathways only one molecule of PEP per molecule of glucose. For this system, it is quite simple to understand how a set of parameters may lead to the trivial (empty) steady state. In the extremely reduced glycolysis system presented in Fig. 4.7 where the diversion arises at rate v_3, only some sets of parameters lead to steady states different from the null steady state, and even then such steady states are not necessarily stable.

Before searching for an optimal model that agrees as well as possible with the experimental data and the prior distributions, it is important to produce a first model with a non-null steady state, even if the fluxes are not very realistic. I will thus now enumerate the steps I use for producing such an initial model.

Ensuring proper flux directions

Every reversible reaction in theory is capable of driving fluxes in either the forward or backward directions. However in my effort to produce a model in agreement with the flux measurements of Haverkorn van Rijsewijk & al. (which defines the conventional forward directions), I assess the agreement using a log-normal distribution, but such a distribution is only defined only for positive values.



Possible non null steady states:

$$PEP^{SS} = \frac{V_2 + k_3K_2 - k_1K_2}{k_1}$$

$$Pyr^{SS} = \frac{k_1K_2 + V_2 + k_1PEP^{SS}}{k_4 + k_4K_2/PEP^{SS}}$$

Linear stability about the steady state:

$$\lambda_1 = -k_1 + k_3 - V_2\frac{1 - \frac{PEP^{SS}}{K_2+PEP^{SS}}}{K_2 + PEP^{SS}}$$

$$\lambda_2 = -k_4$$

Figure 4.7: Simple model of the coupling between the consumption of PEP in upper glycolysis and its regeneration in lower glycolysis. The first reaction consumes one PEP and produces two PEPs, the net PEP yield is thus one. The null concentration for PEP and Pyr is always a steady state solution and its stability depends on the sign of $(-k_1 + k_3 - V_2/K_2)$, negative values corresponding to stability. Non null steady states occur only if the formula for PEP^{SS} leads to positive values, in which case this state is stable if the two eigenvalues of the linearized dynamics about the steady state are negative. From the expression of λ_2 one can see that the stability of the steady state is favored by a small value of k_3 and by a small saturation constant K_2.

It is thus far better to ensure that the models considered have their fluxes of the correct sign. I start by setting the parameters and concentrations at the center of their distributions; in that situation however, the fluxes are not all positive. The first step to fix that consists in shifting the concentrations upstream of a negative flux in order to displace the equilibrium. The standard way to quantify the deviation of a reaction relative to equilibrium is via the ratio Q/k_{eq}. In this quantity, Q is the quotient of the reaction (defined in Eq. 2.2, 4.9) and is a positive quantity. The smaller it is, the more favorable is the reaction, while the value 1 corresponds to equilibrium, and so separates the cases of a forward and backward flux:

$$Q = \frac{\prod_j p_j}{\prod_i s_i} \quad \left(\sum_i S_i \rightleftharpoons \sum_j P_j\right) \tag{4.9}$$

with s_i and p_j being the concentrations of S_i and P_j.

My approach is iterative. I developed a script that tries to improve the fluxes by searching at each step for the reaction having the greatest Q/k_{eq} (if Q/k_{eq} is greater than 1, the reaction occurs in the wrong direction). I then change the concentrations to reduce Q/k_{eq}. The script does the search and updates concentrations until a satisfactory set of concentrations is obtained and produces positive fluxes. I stop the script once the largest Q/k_{eq} ratio has a value of 0.95.

Constructing a steady steady state

The model with all positive fluxes is not at steady state and there is high chance that if the reactions are executed, the final concentrations may go to null values. The simplest way to obtain a steady state with non-zero fluxes in the correct direction is to scale its V_m parameters so that the reference fluxes are obtained for all reactions:

$$V_{m,i}^{new} = \frac{F_i^{ref}}{F_i} V_{m,i}^{old} \tag{4.10}$$

Here $V_{m,i}^{old}$ and $V_{m,i}^{new}$ are the values before and after scaling of the reaction's V_m parameter. F_i and F_i^{ref}

are the instantaneous flux before scaling and the reference flux. The three reactions $Kgdh$, $out_{\alpha Kg}$, and out_{SucCoA} are not associated with any value in the reference (lack of experimental measurements) but they are constrained to be compatible with the published steady state, hence I simply set them manually to appropriate values to obtain a model in the steady state.

Stability of the steady state

The previous procedure leads to a model realization at a steady state, but one has to make sure that this state is stable. Checking that is a simple task, done e.g., by slightly perturbating the concentration of one or many metabolites in the network (a 1% perturbation in the present study), and then simulating the model's dynamics. If the dynamics take one back to the same steady state, one has stability and one has a proper initial model for intiating the optimization process (cf. next paragraph). If on the other hand the simulation gives a new steady state, it means that either the basin of attraction of the constructed steady state is very small or more likely that the steady state is unstable. In any case, if the perturbed steady state does not come back to itself under the dynamics, it goes to another state, in practice also a steady state. I consider this new steady state thereby which by construction will be stable. If it has null concentrations, it is rejected and the whole process of constructing a steady state has to be started over again. In principle such iterations have to be implemented until one obtains a satisfactory stable steady state with fluxes of the proper sign. These iterations rely on the possibility to randomize the construction of a model with a steady state. For example, instead of scaling the V_ms, one may try a random set of V_m and run the dynamics until it reaches a steady state. Fortunately, in my model, the scaled V_m produced a stable steady state so no iterations were necessary.

4.3.3 An algorithm to search for the best model

There are different techniques to search for parameters maximizing the likelihood function. All of them try different sets of parameters in a more or less clever ways and keep the set that resulted in the best score, *i.e.*, the highest log-likelihood. The present model contains 242 parameters, among which some have a prior distribution that varies over several orders of magnitude. There is no hope to run an extensive exploration of the parameter space, and a purely random search will fail to find the best model. The time to generate a random set of parameters, to simulate the dynamics, to find a stable steady state and to compute the score associated with this set of parameters is about 0.3 sec on a core of a personal laptop. On a grid that covers a space between $\mu-\sigma$ and $\mu+\sigma$, μ and σ being the parmeters of the log or log-normal distributions, with a step of 0.01σ for each parameter, it would take 10^{88} years to explore the parameter space.

Fortunately, many algorithms exist to explore parameter sets in a more efficient way. Generally there is no formal way to predict whether an algorithm will be better than another for optimizing a model. Optimization algorithms come in many different types. For my work, I have chosen a genetic algorithm described in annex D.3; such algorithms are quite popular because they tend work well even when using relatively simple implementations. Genetic algorithms are inspired from evolutionary biology where populations are subject to natural selection. After an initial population is generated, the algorithm mates individuals (realisations of the set of parameters in our framework) by pairs, then allows for "mutations" and then performs a selection on these to produce the population of the next generation. The candidate individuals (children of the parents) are selected based on their fitness (log-likelihood in our case) to create a next generation, typically with the same size as the previous generation. The algorithm iterates this generational process which tends to improve the fitness of the population, and the search stops when no further improvement seems possible.

Given the number of parameters in my model and the ranges (orders of magnitudes) for the priors of these parameters, I found that I could not use already existing genetic algorithms and had to instead develop my own program, inspired of course by uses of practitioners of genetic algorithms.

Initialization of the search population A first generation of size $N = 10$ is obtained by mutating the initial set of parameters corresponding to the stable steady state developed in the previous section. The

mutation rate is 5% which means that $0.05 \times 242 \approx 12$ parameters are perturbed according to this change:

$$p^{new} = p^{old} \times (1 + s\sigma\xi) \tag{4.11}$$

The quantities p^{new}, p^{old} are respectively the new and the old value of the parameter when it is distributed normally or the log of the parameter when it is distributed log-normally. s is the strength of mutation, σ is the parameter of the normal or log-normal distribution, and ξ is a random variable taken in a Gaussian distribution of mean 0 and variance 1. The value of s is inially set to 0.01.

The size of the population is $N = 10$ but other choices are possible. Before being included in the initial parameter population, a parameter set must lead to a non-null stable steady state. Given that a model has a proper steady state, its score is computed.

Note that in Fig. 4.7 we see that a model may be extremely sensitive to parameter perturbation which can change the stability of a steady state and the newly generated parameters may produced null fluxes, which leads to rejection of the model. Beyond the concentrations and fluxes of the model, I measure relaxation and mean passage times. If any of those times are larger than 5000 sec, the parameter set is also rejected.

Producing children by cloning and mating parents To produce the $(i+1)th$ generation, I first increase the population size by copying (cloning) the N individuals, creating N children. Then I also generate 10 further individuals by mating amongst parents. The mating is done by randomly selecting two parents in the generation i, all parents having the same probability to be picked. The child is built with half of its parameters coming from the first parent and the other half coming from the second parent.

Mutations The $2N$ children undergo mutations with a rate of 5%. The parameters selected for mutation are transformed according to Eq. 4.11.

Selecting the individuals for the next generation The new steady state is calculated for each child, and then the corresponding score is computed. Children that were not able to produce a satisfactory steady state (null fluxes, too long transit times) are rejected. After this rejection, it is possible that one is left with fewer than N children; if so, a new population of children is produced again and the satisfactory ones are added to the former ones. The production of a new generation of children continues as long as the total number of satisfactory ones is smaller than N. Once the number of satisfactory children is large enough, the N children with the highest scores are selected to form the generation $i+1$.

Keeping track of the best model found so far From a practical point of view, the best model found is always kept in memory, it is changed only when a better model is produced. Every 500 generations, the algorithm takes a snapshot of the best model and saves it in a file. If the best model has not improved since the last printing, the algorithm stops as presumably further searches are likely to be futile.

From rough to refined exploration The printing steps is also a moment where the parameter perturbation amplitude, s, is modified. At each generation the algorithm computes the number of parameter sets that has a greater score than the mean score of the parents. If on averge across the 500 hundred generations this number is lower than 20%, s is multiplied by 2 otherwise it is divided by 2. This techniques allows the fine tuning of the perturbation size as one gets closer to optimality. To prevent the algorithm from encountering numerical problems, I impose the lower bound of 10^{-5} on s.

4.4 Estimation of the confidence interval for the parameters

It is common that the result obtain from optimization is not sensitive to parameter variations. It is hard to tell a parameter has a certain value when one may vary it without impacting the score. I propose here a Bayesian method to estimate the confidence one can have on an optimal parameter value. The global idea is to reconstruct the parameter's likelihood distribution. The standard method to do so is to use a Markov chain Monte Carlo (mcmc) algorithm. It is a class of algorithms that perform a random walk

in the parameter space and save the points it visits, the more likely a points is the more it is saved. To ensure that the algorithm visits the parameter points with the correct ratio, the transition between two sets of parameters $\vec{p_1}$ and $\vec{p_2}$ is defined as

$$\pi_{1\to 2}\mathscr{L}(\vec{p_1}) = \pi_{2\to 1}\mathscr{L}(\vec{p_2}) \tag{4.12}$$

where $\pi_{1\to 2}$ (resp. $\pi_{2\to 1}$) is the transition probability from $\vec{p_1}$ to $\vec{p_2}$ (resp. $\vec{p_2}$ to $\vec{p_1}$). This equation is called the detailed balance. Note the analogy with the mass action formalism where the concentration stands for the probability of a state. The task of the user is to define an algorithm that can explore the parameter space with appropriate transition probabilities. To do so I used the Metropolis-Hastings [33] algorithm that decomposes a transition into two steps:

- Select randomly a point $\vec{p}_{new}$. Practically, 6 parameters are drawn uniformly from the $\vec{p}_{new}$ and are transformed according to Eq. 4.11 with s taking only to values with equal chance ± 0.01.
- The transition from $\vec{p}_{old}$ to $\vec{p}_{new}$ is calculated according to

$$\pi_{old\to new} = \begin{cases} \mathscr{L}(\vec{p_{new}})/\mathscr{L}(\vec{p_{old}}) & \mathscr{L}(\vec{p_{new}}) < \mathscr{L}(\vec{p_{old}}) \\ 1 & \mathscr{L}(\vec{p_{new}}) > \mathscr{L}(\vec{p_{old}}) \end{cases}$$

  The transition probability verifies automatically the detailed balance Eq.4.12.

If the algorithm saves a point for every perturbation, they will all be closely related since not all the parameters have been perturbed yet. Instead, I wait 250 perturbations before saving my data point so that y parameter is perturbed more than six times on average. The final statistical distributions for parameters are built from 10^5 saved parameter points. A more extended description of the algorithm is made in D.4.

5

Analysis of the model

In this chapter I analyze different aspects of my optimized model, *i.e.*, the model obtained after the parameter fitting process. There are several possibilities for performing such an optimization, in particular I used different choices for the constraints, penalizing or not the parameter values via prior distributions, including or not the transit and relaxation times, etc. The impact of these choices on the overall quality of the fit is described. I use this chapter also to list several applications of such a kinetic model, for instance with respect to parameter inference or for metabolic engineering.

5.1 Result of the model optimization

The optimization of the model relies on (i) estimations of parameter values in the literature for specific reactions, (ii) global distributions of parameters of a given type (used in priors), and (iii) systemic experimental measurements (concentrations, fluxes, etc.). The detailed description of how the constraints are implemented in my optimization process was provided made in chapter. 4. The overall likelyhood of a model's parameters is decomposed into four components:

C The probability that the model leads to the given concentration data

F The probability that the model leads to the given flux data

P The probability of the model's parameters given the prior distributions

T The probability that the model leads to the experimental or prior values for characteristic times (the relaxation time for concentration perturbations, the mean passing time for Pyr, DHAP, Ace, and AcCoA).

These four components of the likelihood function are represented by the letters C,F,P and T. To evaluate the importance of these different components, in particular of the P and T components, I performed different optimisations for various choices and compared the results. Note that the time-independent concentration (e.g., ATP) are classified as parameters and contribute to the P component of the score, not to the C component.

A typical score improvement curve is displayed in Fig. 5.1. The score improvement is initially very rapid and then progressivelly slows down as the search becomes fine tuned as it converges. The result of the optimisation for the four constraints CF, CFP, CFT and CFPT are listed in the Annex C.

5.2 Optimisation using only fluxes and concentrations

As a first choice of optimisatisation, denoted CF, I impose in the likelihood only the terms associated with C and F. This enforces agreement of the model with the curated experimental data from Bennett's concentrations [7] and Haverkorn van Rijsewijk's fluxes [34]. The parameters are allowed to vary without impacting directly the score (I include no priors) and only the deviations between the model's steady state predictions and experimental measurements matter, corresponding to the log-likelihood score

$$\log \mathscr{L}(\vec{p}) = -\sum_i \frac{(\log F_i - \mu_i)^2}{2\sigma_i^2} - \sum_j \frac{(\log C_j - \mu_j)^2}{2\sigma_j^2}$$

with the notation for parameters being the same as in Tab. 4.3.

The optimization I ran produced a final best log-likelihood score of -3.63. In direct analogy with what happens for a chi square (modulo a change of sign and prefactor), the score will decrease by $1/2$ if the value entering a term of the score (say one particular concentration) goes from its optimum value (the central value) to a value one standard deviation away. The more negative the model's score, the more it has difficulty reproducing the target behavior, *i.e.*, satisfying all the constraints simultaneously. In this CF optimisation, 22 concentrations and 42 fluxes are constrained, a score of -3.63 means that the algorithm managed to get very close the reference concentrations and fluxes. This is no doubt the case because of the many parameters of the model, but since all these parameters do not act independently, it

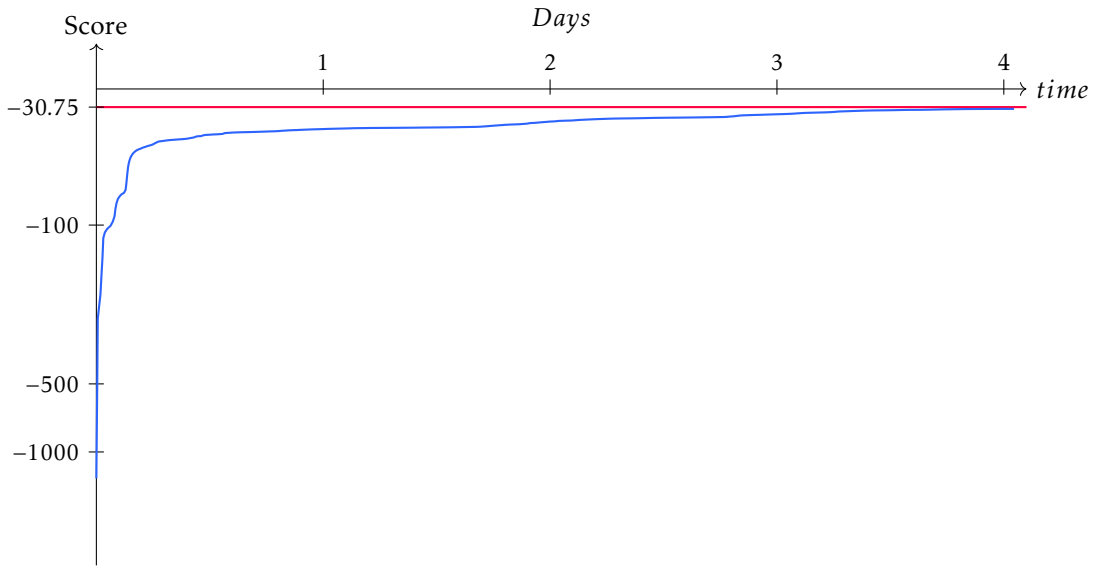


Figure 5.1: Score (log-likelihood) of the best found model as a function of computation time when using the CFPT constraints. After 4 days, the run stopped at a best score of -30.75.

is not possible to guess a priori to what degree agreement is possible as there are undoubtedly a number of trade-offs included. Of course this is helped by the fact that there is no prior for these parameters: the optimization proceeded without the term P, allowing arbitrary flexibility in the choices of the parameter values.

Given the CF optimisation, one may ask whether the corresponding optimal model has anomalous parameter values in the sense of being outside the prior distribution (component P of the likelihood function but not used during the optimization). One can similarly ask what kind of characteristic times are obtained (component T). The values of the scores for these quantities are provided in Tab. 5.1. First, I find that the parameters do not lie outside of their prior distributions since the P component of the score is −95 for 241 parameters: effectively, only 24 inferred parameters lie further than one standard deviation from the mean of their prior distribution. Second, I find a different result for the T component. Indeed, the optimisation ignoring the T part of the score does not lead to satisfactory characteristic times since the times of mean passage for the tracers are $T_{Pyr} = 58.6\ s$, $T_{DHAP} = 59\ s$, $T_{AcCoA} = 58.7\ s$, and $T_{Ace} = 58.7\ s$. These values are to be compared with the (unused) references of 12.9 s, 33.2 s, 8.96 s, and 9.06 s. However, the relaxation time of the model, 19 s, compares reasonably with the 13 s value of the reference and thus is more acceptable. Lastly, I looked into the predictions from the CP optimisation for the predicted values of the formation free energies. If they hardly moved it is because their value impacts strongly the behavior of the network and one can assume that their theoretical value (from eQuilibrator) is in fact very close to the true value.

Score component	total	C	F	P	T
value	−109.3	−1.37	−2.26	−95	−10.7

Table 5.1: Values in the CP optimized model of the different score components.

The optimization of the model was implemented using a genetic algorithm. That choice was made because such algorithms are quite robust and able to avoid local minima. Nevertheless they are heuristic in the sense that they do not guaranty that the global maximum will be found. Their good behavior in practice is generally interpreted by saying that under the search procedure the global optimum has a large basin of attraction. The final score is remarkably close to its maximum value (0), and it is quite likely in my opinion that the global optimum has been found.

Since my approach is based on a Likelihood function, it is possible to not only provide an (optimal)

model with specific parameter values, it is also possible to evaluate the confidence on those parameter values. Some parameters may have a very little impact on the global score, namely they can vary over a relatively wide range without significantly degrading the model's score. In such a situation, one expects them to be quite poorly determined. To compute the confidence intervals, I followed the (Bayesian) framework presented in chapter 4: I sampled an ensemble of models using as measure the likelihood. The Bayesian interpretation of the models in this ensemble show that the distribution obtained during the sampling provides the a posteriori distribution of the parameter values, from which it is easy to extract their confidence intervals.

The free energies of formation of the different metabolites are very well defined. As mentioned above, a variation in these values changes the equilibrium constants and the overall metabolic network behavior is quite sensitive to these values. To drive this point home, consider for instance the Pgi reaction G6P $\rightarrow$ F6P. The free energies of formation of G6P and F6P are $-130.34\ KJ$ and $-130.09\ KJ$, thus the reaction energy is $-130.09 + 130.34 = 0.25\ kJ$. A 1% change of a free energy of formation (eg $\Delta_f G^{F6P} \times 1.01$) leads to a large relative change in the reactional energy :$(1.3 + 0.25)/0.25 = 600\%$! A large perturbation of that scale destabilizes all the fluxes in the network. For this reason, there is a large penalty for the score whenever the free energies of formation are changed significantly, thus the initial claim that these values are in fact obtained to very high precision.

5.3 Optimisation imposing CFP, CFPT and CFT constraints

In the CFP ensemble which adds the P terms in the likelihood when compared to the CF case, those added terms correspond to a penalty for variations in V_ms, K^Ms, and $\Delta_f G$s. The quality of the model now depends on the agreement of the model's fluxes and concentrations with the reference data but also on the positions of the parameters in their reference distributions. Performing the optimization run, the optimal score found was equal to -58.3 which is, as before, much smaller than the number of constrained quantities (305). But it also shows that the system prefers particular values of the parameters that are not the ones provided by the priors. Let me note furthermore that when running the optimisation several times with different initial values for the parameters, I was led to the same set of final values for the parameters.

I then ran the optimization using the CFPT terms in the likelihood. This led to a best score of -30.75, in spite of having only added constraints to the CFP ensemble. Indeed, this CFPT approach adds terms for the agreement of the charactistic times with the reference relaxation times and mean passage times. This result may seem paradoxical but it means that the CFP optimisation runs did not provide the global maximum, and in fact got stuck in the samle local minimum. This result shows that the addition of the T type constraints actually helped the genetic algorithm find better models. Another feature is that the characteristic times associated with the CFP optimum are much greater than those of the CFPT. The two sets of times are presented in the two first rows of Tab. 5.2. Not surprisingly, the times produced in the CFPT ensemble are systematically lower than those in the CFP ensemble. Nevertheless, the mean passage times obtained are definitely larger than the reference times while the value predicted for the relaxation time is only a bit above the target value. For completeness, Tab. 5.3 lists the fluxes and the concentrations obtained imposing the CFPT constraints.

	T_{Pyr}	T_{DHAP}	T_{AcCoA}	T_{Ace}	τ
CFP_1 (s)	121	55	121	121	56
CFPT (s)	51	50.9	51	51	15.9
CFP_2 (s)	58.2	58.4	58.2	58.3	19.4
CFT(s)	50.6	50.7	50.6	50.6	16

Table 5.2: Top two lines: characteristic times produced by the optimization runs using the CFP and the CFPT ensembles. Bottom two lines: the same times but for the CFP and CFT ensembles when taking as initial parameter values the optimal model produced by the CFPT run.

The poor job done by the optimization based on the CFP ensemble, both in terms of score and the

Metabolite	Optimal concentration	Constraint
PEP	0.16	0.18
G6P	7.81	8.75
F6P	0.13	
Pyr	0.43	∅
FbP	12.19	15.20
GAP	0.03	0.37
DHAP	0.36	
BPG	0.06	∅
PGA3	1.41	1.54
PGA2	0.12	
GL6P	0.90	1.04
PGn	3.70	3.77
Ru5P	0.23	1.32
R5P	0.30	
X5P	0.68	
E4P	0.07	∅
S7P	0.47	∅
KDPG	0.25	∅
AcCoA	0.58	0.61
Ace	0.45	∅
OAA	0.01	∅
Icit	0.06	1.96
Cit	1.77	
aKg	0.34	0.44
SucCoA	0.17	0.23
Suc	0.44	0.57
Fum	0.53	0.12
Mal	1.67	1.68

Reaction	Optimal rate	Constraint
PTS	1.05	1.05
Pgi	0.76	0.73
Pfk	0.83	0.84
Aldo	0.83	0.84
Tis	0.83	0.83
Gdh	1.78	1.80
Pgk	1.78	1.80
Pgm	1.64	1.68
Eno	1.64	1.68
Pk	0.27	0.24
Zwf	0.28	0.31
Pgl	0.28	0.31
Gnd	0.19	0.21
Rpi	0.11	0.11
Rpe	0.07	0.10
TktA	0.02	0.04
TktB	0.05	0.07
Tal	0.05	0.07
Pdh	1.16	1.18
Acs	0.72	0.71
Ppc	0.28	0.31
Cs	0.28	0.29
Acn	0.28	0.29
Icdh	0.28	0.29
Kgdh	0.19	∅
Stk	0.09	0.17
Sdh	0.09	0.17
FumA	0.09	0.17
Mdh	0.07	0.10
Mae	0.02	0.06
Edd	0.10	0.10
Eda	0.10	0.10
out_{G6P}	0.01	0.01
out_{DHAP}	0.00	0.01
out_{R5P}	0.07	0.04
out_{E4P}	0.03	0.03
out_{PGA3}	0.14	0.12
out_{PEP}	0.04	0.07
out_{Pyr}	0.28	0.27
out_{AcCoA}	0.16	0.19
out_{Ace}	0.72	0.71
out_{OAA}	0.08	0.13
out_{aKg}	0.09	∅
out_{SucCoA}	0.10	∅

Table 5.3: Comparison of values predicted by the optimized model using CFPT constraints to the experimental values. The left table concerns the concentrations of metabolites or pools. The right table concerns the fluxes through reactions. The ∅ symbol means that there was no experimental measurement. The units are mM for the concentrations and $mM\ l^{-1}$ for the fluxes.

too long characteristic times found, can *a posteriori* be justified via the high saturation values produced. I prefer to consider saturation via a broader perspective than its usual definition: I take its measure as reduction in the flux due to the different regulatory effects, including activation and inhibition as well as the effects of the denominator in convenience kinetics. Based on the general formula

$$v = V_m \left(\prod_i s_i - \frac{\prod_j p_j}{k_{eq}} \right) sat(\vec{c}) \tag{5.1}$$

I consider the saturation to be $sat(\vec{c})$. This factor may depend, in general, on the concentrations of various metabolites in the network. Although $sat(\vec{c})$ may encompass different regulatory effects, it mainly accounts for substrate and product saturation in the considered reaction. In its simplest form, the saturation acts on the reaction as in the MMH rate law (cf Eq. 2.2). Namely, when the concentration of substrate is much larger than the scale K^M, the reaction occurs in the saturated regime where even large changes of substrate concentration have little effect on the reaction rate. It turns out that the optimization algorithm (and presumably this is not specific to our genetic algorithm) naturally gets driven to this type of regime because then it can adjust the concentrations and the V_m independently. As support for this hypothesis, note that the saturations are less important (larger $sat(\vec{c})$) for the optimization in the CFPT ensemble than in the CFP ensemble. Interestingly, I ran the CFT optimization but starting from the model produced by the CFTP ensemble to find that the CFT optimisation kept improving its score until reaching a value of -20.4, with a relaxation time presented in the third row of Tab. 5.2.

The trend towards saturation found in the CFP ensemble naturally leads to long characteristic times, simply because when a reaction is saturated it can no longer react to changes and thus the characteristic times are long. We believe that this bias towards saturation is a general phenomenon, independent of the actual algorithm doing the optimization search. I have found it in other published models. For example, one of the largest kinetic models that exists for yeast [69] (available on the BioModels database) has a relaxation time of $449\,h$, which seems so large as to discredit the model. Those authors used a model development having points in common with what I used here but they did not account for any characteristic time. This shows the importance of imposing time constraints for the optimisation. Coming back to my work, I seeked to reduce the remaining discrepancy between my models and experiments for the mean passage times. One possible cause of this discrepancy is the use of priors on many of my parameters. I thus implemented the CFT optimisation (no inclusion of penalties from P terms) and used as starting point the optimum model produced in CFPT. Surprisingly, the discrepancy remained, the mean passage times were still significantly too long, and very little difference was found compared to the CFPT ensemble (cf. fourth row of Tab. 5.2).

The CF optimisation seem less affected by the bias towards saturation than the CFP. Note that the CF optimisation is underconstrained and that it has a larger flexibility for the choice of V_m since these parameters do not have lie within a reference distribution. I find that almost every V_m is larger for the CF optimum than for the CFP optimum. Note that the CF and CFT optima have respectively 24 and 32 parameters that are further than one standard deviation away from the reference distributions. For the CFP and CFPT optima, the corresponding numbers are far lower as they should, being respectively 3 and 6. The 3 "outlying" parameters of the CFPT correspond to the concentrations of Phosphate, NADH and NADP that were taken according to Bennett's published values; these may be a bit different in the conditions that led to the optimised fluxes. Adding the constraint of having the parameters lie in the prior distribution narrows the allowed parameter space, and having many parameters, especially K^M, near the center of their reference distribution indicates that new constraints are necessary to obtain the same distribution width as for the dispersion of actual data. Nonetheless the confidence intervals are small for the parameters shown in the annex C; for a set of constraints, the parameter landscape (likelihood function) is relatively sharp about the optimum. Such a parameter landscape would resemble more the blue landscape in Fig. 5.2 than the red one.

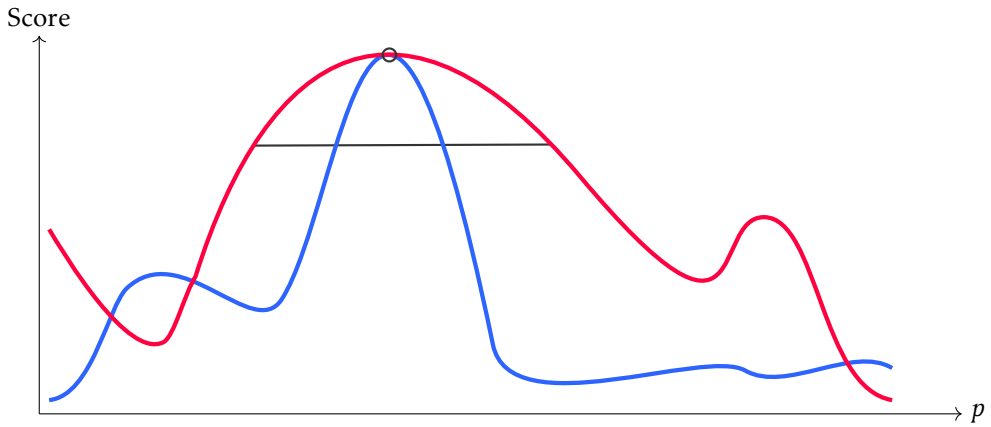


Figure 5.2: Two likelihood landscapes schematically represented for a one parameter system. Both share the same optimum symbolised by a circle. The confidence interval corresponds roughly to a span over which the score remains close to the maximum value. The vertical line sets a visual boundary for the significance but in practice the confidence interval is evaluated by building the histogram of values based using a MCMC sampling and taking the range where a given fraction of the events lie. In the example, the blue landscape has a smaller confidence interval so the optimum value will be more precise than when using the red landscape.

5.4 Quotients of the reactions in the model

Early in my thesis, I chose to work within a kinetic framework of metabolism that was automatically consistent with thermodynamics. Even a reaction strongly favoured thermodynamically may have a net flux in the backward direction when it is set in conditions with high concentrations of products and low concentrations of substrates. Therefore the quotient of reaction, defined as in Eq. 2.2, 4.9, is a better criterion to test the degree of reversibility of a reaction in practice because it includes the physiological conditions of reactant concentrations. Q is a positive quantity and the direction of the flux is given by its value according to

Recall that the reaction free energy is proportional to $\log Q$, thus it releases energy – is favourable – for values of Q smaller than 1. The figure Fig.5.3 shows the quotient of reaction for the set of all reactions in my model. Some clear trends are visible. The glycolysis pathway is thermodynamically close to equilibrium (highly reversible). This property is necessary to allow gluconeogenesis (production of glucose from other substrates like Pyr) and growth on other substrates than glucose. The Entner-Doutoroff pathway is more out of equilibrium. It is known that it is prefered under low energy availability because it requires less enzymes synthesis. Lastly, the reaction Mdh in the TCA pathway is known for not being favoured thermodynamically and needs its product, OAA, to be in very small amounts to proceed in the forward direction.

5.5 Control coefficients in the network

In biotechnological applications involving engineering biochemical pathways, one wants to find what is the limiting step in a metabolic network, step which prevents a faster production of the compound of interest. In reality, searching for one unique limiting step is a naive expectation and that will be

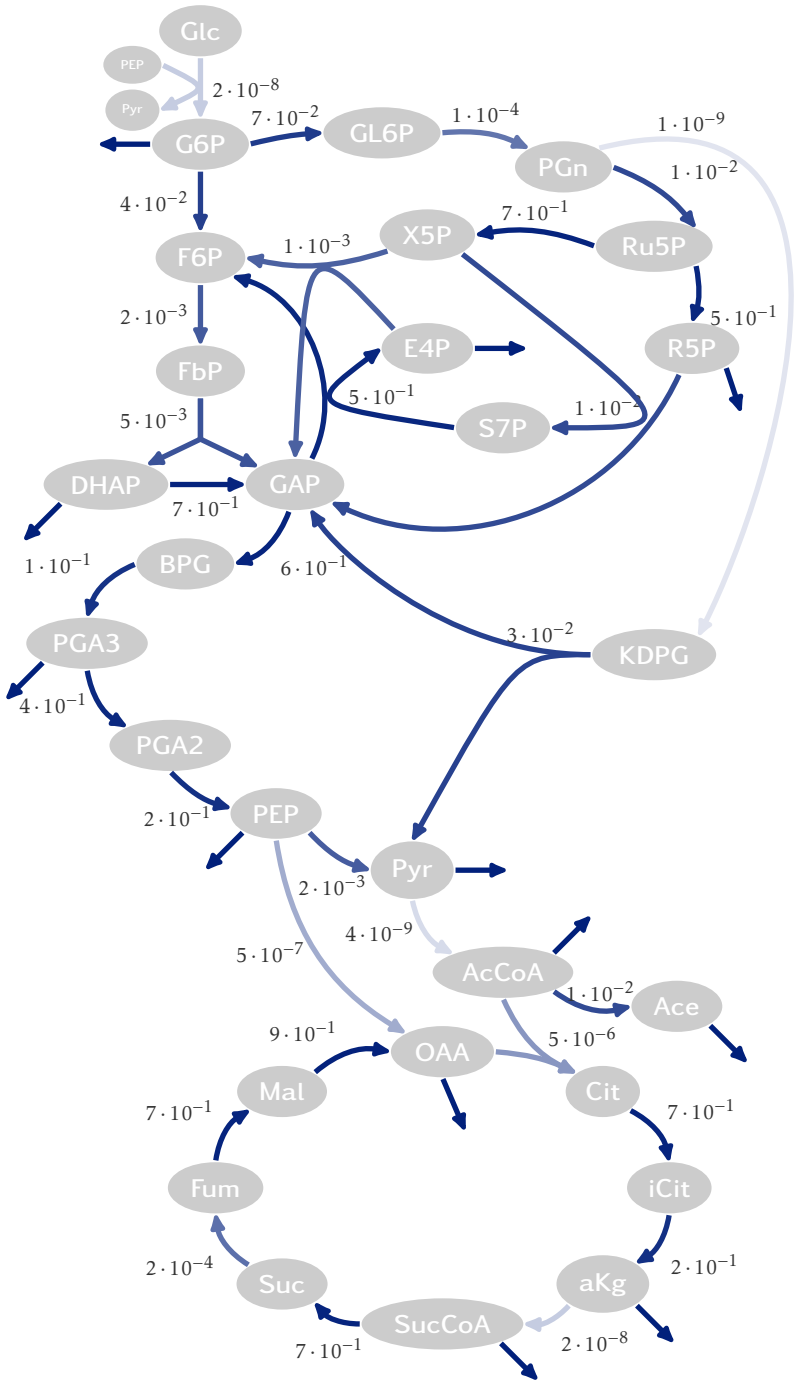


Figure 5.3: Quotients of the reactions in the network. Darker blue arrows correspond to quotients of reactions close to 1. Light colored arrows stand for relatively irreversible reactions. The direction of the arrows idicate the conventional forward direction.

successful only in simple networks where all the metabolites but one are in excess. For steady-state conditions, the notion of "limiting" metabolite or enzyme tends to be irrelevant because the fluxes, which depend on the abundances of enzymes, form an interconnected network. At steady state, increasing one rate by adding more enzymes potentially has an impact on all the network fluxes, thus every enzyme exerts an effect on all fluxes. Under such conditions, a more systematic approach is necessary to quantify how the abundance of each enzyme in the network may affect the flux one is trying to increase. In 1973, Kascer and Burns proposed a mathematical formalism [42] to describe the effect of each network enzyme on any flux of interest. The *control* of a flux is shared between the enzymes and each enzyme can contribute to a certain portion of the total control, contribution which is quantified by a control coefficient. The *control coefficient* for an enzyme accounts for the relative change in the flux of interest after a perturbation of the enzyme concentration. In the example that aims at improving out_{DHAP}, the control coefficient for a generic enzyme E of concentration e is given by:

$$C_E^{\text{out}_{DHAP}} = \frac{d \log v_{\text{out}_{DHAP}}}{d \log e} \tag{5.2}$$

where $v_{\text{out}_{DHAP}}$ represents the steady state flux of the reaction driving DHAP towards biomass (exiting the system). The total control on a given flux obeys to a so called "sum rule". Summing the control coefficient of every enzyme leads to:

$$\sum_E C_E^{\text{out}_{DHAP}} = 1 \tag{5.3}$$

Because the k_{cat} parameters are fixed, we have

$$d \log V_m = d \log e + d \log k_{cat} = d \log e$$

and Eq. 5.2 can be re-written as

$$C_E^{\text{out}_{DHAP}} = C_{V_m}^{\text{out}_{DHAP}} = \frac{d \log v_{\text{out}_{DHAP}}}{d \log V_m}$$

This new formulation is more convenient for evaluating the control coefficients in the network since V_m is an explicit parameter of the reactions. The Fig. 5.4 displays the control of every enzyme in the network on the out_{DHAP} flux towards consumption of that metabolite (exiting from the CCM). As it had to be, the coefficients verify the sum rule, Eq. 5.3. The control coefficients are not necessarily all positive, a positive coefficient means that increasing the enzyme concentration (or the reaction's V_m) is beneficial for increasing out_{DHAP} whereas a negative control coefficient means that the out_{DHAP}'s flux is negatively affected when increasing the reaction's enzyme concentration.

In the perspective of the RESET project, the control coefficients may be very informative, they indicate whether a decrease in a flux's capacity impacts the production of glycerol. It is reasonable to start from the hypothesis that the reduction of some of the exits from the system leads to a global increase of concentrations and enhances the exiting fluxes that are not involved much in feeding into biomass production. In other words, by shutting down the exits toward amino acids and nucleotides, the other exits should be favored. This argues for the control coefficient for the such "other" exits to be negative for the reactions leading to biomass production. However the control coefficients show a more complex behaviour. The hypothesis is true for the exits that are inhibited in the PPP (out_{E4P}, out_{R5P}) because when these exits are active, they favour fluxes in a pathway parallel to the CCM and drive carbon away from DHAP. However the exits at the bottom of the CCM (TCA, PEP, Pyr) have a positive control coefficient. The explanation for this phenomenon is that the out_{DHAP} alone is not enough to sustain the strength of glycolysis and if all the other fluxes at the bottom are stopped, the whole glycolysis is inhibited too, a consequence which is not profitable for the production of glycerol. Therefore the control coefficients do not allow to conclude about whether the RESET project can be successful at improving out_{DHAP} by stopping or reducing uniformly the exits toward amino acids and nucleotides. The more thorough test that I will now describe in the next chapter are necessary to conclude on what happens when the fluxes towards amino acids and nucleotide are reduced.

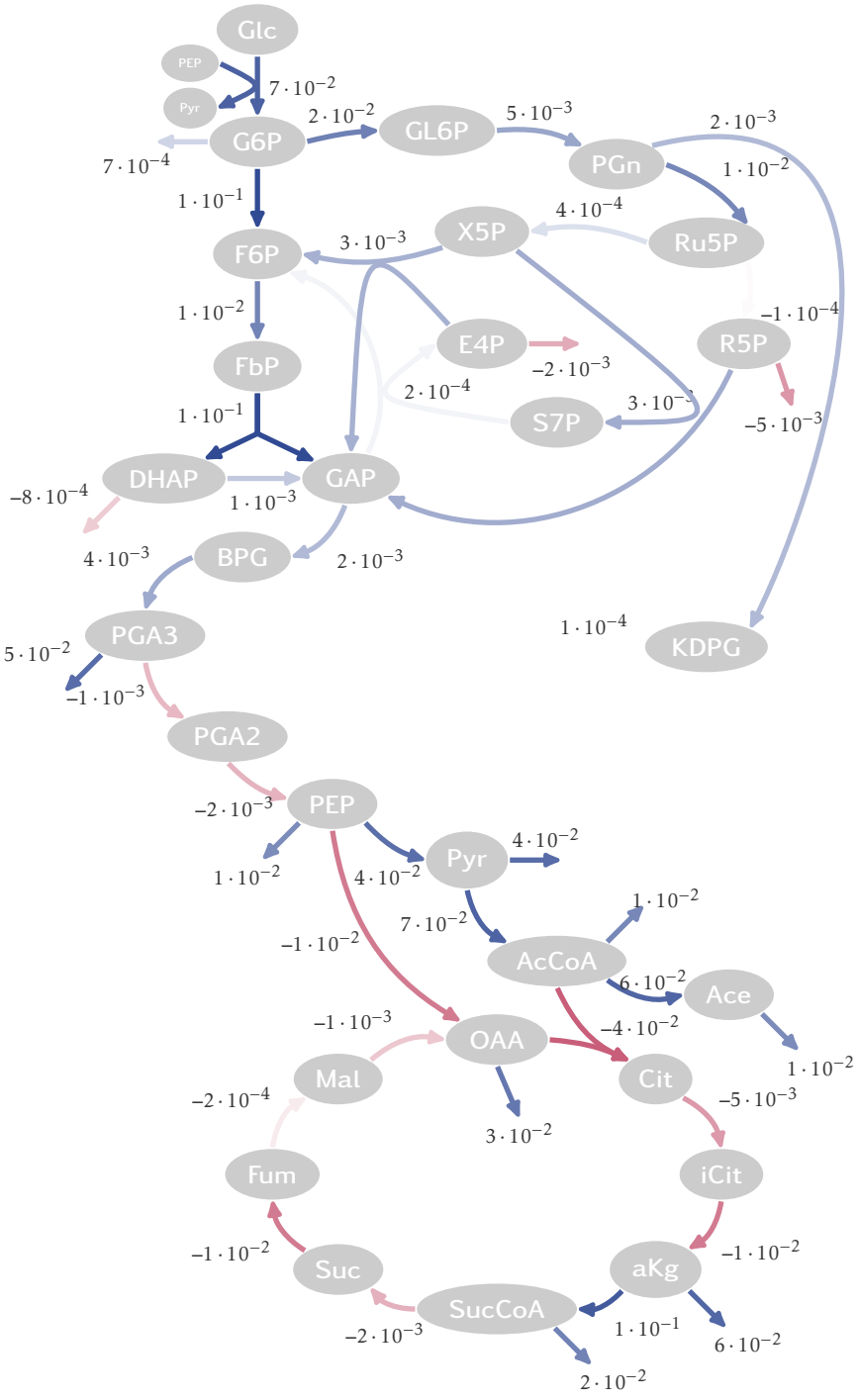


Figure 5.4: Control coefficients for the exiting flux out_{DHAP} (that leads to glycerol), for each of the enzymes in the network. Red arrows show the enzymes with a negative control coefficient and the blue arrows show the enzymes with a positive control coefficient. The darker colours correspond to a large absolute value of the control coefficients.

6

Use of the kinetic model to test the RESET strategy

One major goal of this thesis was to create the computational tools to test the RESET scenario *in silico*, namely finding out whether turning off the *gene expression machinery* (GEM) can lead to modified fluxes in the metabolism so that glycerol production is improved. In the chapters so far, I have presented the methods that allowed me to build a full kinetic model of *E. coli* 's central carbon metabolism which agrees well with physiological and other measurements taken from the literature. In this chapter, I exploit my optimized kinetic model, using it to perform the kind of manipulations that are relevant for the RESET project. Ultimately, my kinetic model will be coupled with a full scale GEM model. In the mean time, I show here a few test performed directly on my model by using various proxies of manipulations of the gene expression machinery, expecting that these can provide good indications of whether the RESET strategy may be successful.

6.1 Shutting down consumption of precursors

The RESET strategy relies on the non consumption of amino acids or other biomass building blocks to inhibit the corresponding biosynthesis pathways. Such inhibition is ubiquitous in biochemistry, and corresponds to the quite simple operating logic of the type "just in time" that has become common in the merchant world: provide if there is a demand, don't provide if there is no need for the product. In terms of the mechanistic implementation of such logic in biochemistry, many cases are known, for instance in AA biosynthesis, where the final product can bind to one of the upstream enzymes of the pathway and thereby inhibit its action. In effect, each fundamental building block used in biomass production via polymerisation (amino acid, nucleotide, fatty acid) can slow down the biosynthesis pipeline that is responsible for its production. In *E. coli* , the ratio between amino acids (polymerized or not) and nucleotides (polymerized or not) is about 8 : 1 [57], so for the rest of this study, the reader may worry principally about AA since their contribution dominates that of nucleotides. Nucleotides and amino acids are formed from eight precursors only: PEP, Pyr, αKg, PGA3, R5P, E4P, AcCoA, and OAA. Therefore a first test of the RESET scenario is to inhibit partially or completely the outgoing fluxes of these precursors. My model of the CCM involves effective reactions which absorb these metabolites as a way to model their consumption by the biosynthetic pathways, so it is enough to manipulate those reactions.

In practice, the parameters V_m of these effective reactions, which remove these metabolites from the system, can be modulated as a proxy for reducing the fluxes through the biosynthetic pathways of AA and nucleotides. I do this by rescaling the V_m values:

$$V_m^{new} = (1-\alpha) \times V_m^{old} \tag{6.1}$$

where α is a proxy for the intensity of the RESET-like perturbation; a value of 0 stands for the optimised model and a value of 1 corresponds to a complete inhibition of the biosynthesis pathways. To make the test as simple as possible, I used the same value of α for all eight precursors. Then I determined the new steady state in this modified (perturbed) model and measured the yield in glycerol, defined as the flux of glycerol production divided by the flux of glucose uptake:

$$\text{yield} = \frac{v_{out_{DHAP}}}{v_{PTS}} \tag{6.2}$$

Naturally the yield in such a perturbed situation depends on α. Tab. 6.1 provides some yield values to illustrate the qualitative behavior as a function of α. There is a clear increase in the yield values as the

α	0	0.25	0.5	0.75	1
yield	4.15×10^{-3}	4.21×10^{-3}	4.29×10^{-3}	4.49×10^{-3}	5.64×10^{-3}

Table 6.1: Yields predicted by the *in silico* modulation of the consumption of the biomass precursors when operating in the steady state regime. α is the proportion by which the fluxes are shut down.

precursors see their consumption go down, even if the magnitude of the effect is modest. As expected,

the carbon that is not used in the biosynthesis pathways is recycled in the various other exits from the CCM. The non utilization of the biomass precursors is propagated to the other consumption pathways. Along with that rearrangement of fluxes, there is a global increase of the concentrations of metabolites in the model. One consequence of this increase is to reduce net carbon uptake. In fact, most of the yield increase can be traced to the reduction of the denominator in the formula for the yield. For instance, going from $\alpha = 0$ to $\alpha = 1$ reduces the flux of PTS from 1.05 to 0.79 $mM\,l^{-1}$ while the out_{DHAP} flux goes from 4.36×10^{-3} to 3.93×10^{-3} $mM\,l^{-1}$. My modeling of the RESET-like perturbation thus has the effect of increasing yield but also of reducing flux.

6.2 Taking into account the separate contributions of each precursor to the bio-blocks pools

The procedure for modeling RESET-like perturbations described in the previous section is simple but neglects the fact that all precursors do not contribute identically to the bio-blocks pools. Specifically, not all the flux consuming a given precursor goes to the different pools (amino acids, nucleotides, fatty acids) in the same proportions. Since in the previous section the effects on yield were small, it is wise to reconsider the perturbation modeling to better take into account the different characteristics of each precursor. I do this by separating the consumption flux into different parts: one part for the AA biosynthesis, one part for the nucleotide biosynthesis, and one part for the rest. For example, in the case of OAA, the decomposition using known proportions towards the different end products is written as

$$v_{out_{OAA}} = \left((1 - \gamma_{Nucl} - \gamma_{AA}) + \gamma_{Nucl} \frac{F_{Nucl}}{F^0_{Nucl}} + \gamma_{AA} \frac{F_{AA}}{F^0_{AA}} \right) V_m \frac{C_{OAA}}{K^M_{OAA} + C_{OAA}} \tag{6.3}$$

where γ_{Nucl} and γ_{AA} are the molar proportions of OAA that contribute to the pool of nucleotides and to the pool of amino acids. These proportions are computed from the stoichiometry matrix that associates the precursors to the different biomass pools and the molar fractions of the amino acids in the organism's biomass formula. The biomass components contain amino acids and nucleotides but also other important cell components such as fatty acids, peptoglycans, etc. [57].

In the expression of Eq. 6.3, it is assumed that the cell produces the amino acids and nucleotides in optimal proportions to serve the polymerisation reactions responsible for the biomass formula. Let F_{Nucl} be the consumption flux of the nucleotide pool; it must be equal to the flux through the nucleotide biosynthesis pathways. A reduction of F^0_{Nucl} reduces proportionally the nucleotide polymerization flux and thus that part of the flux out of precursors dedicated to synthesis of nucleotides. Note that when the fluxes of polymerisation are unchanged from their original value, the expression of Eq. 6.3 is identical to Eq. 2.5.

In the development of the modified reaction rate equations associated with a RESET-like perturbation, we must take into account the change in transcription and translation rates of the GEM. This leads us to define the ratio F_{Nucl}/F^0_{Nucl} (cf. the definitions above) and the analogous ration for AA, F_{AA}/F^0_{AA}. In general, and in the RESET project in particular, these two fractions need not be equal. Fig. 6.1 presents the behavior of glycerol yield in the $(F_{AA}/F^0_{AA}, F_{Nucl}/F^0_{Nucl})$ state space. (In fact, what is displayed is the change in yield when applying the perturbation.) Interestingly, all the values are positive which means that every perturbation of the GEM leads to a model with a higher glycerol yield.

These results provide the same qualitative picture as the more crude approach to perturbing the system described in the previous section. In particular, the fluxes toward glycerol are small and remain so under RESET-like perturbations.

6.3 Kinetics of the model after the GEM arrest

So far, I showed that the model allowed me to have an increase of the glycerol yield by applying a RESET-like perturbation when comparing steady-state properties. However, the steady state is not necessarily reached quickly because the even if the gene expression machinery is no longer renewed, there is a

relaxation associated with the decay of the different components. Plus it is not clear what are reasonable values for F_{AA}/F^0_{AA} and F_{Nucl}/F^0_{Nucl} in typical conditions. Until my metabolic model is coupled to a realistic mathematical description of the gene expression machinery, I have used a hands-on approach for representing the GEM via an extremely simple set of equations. These equations provide a coarse-grained summary of the time dynamics of the bulk abundances of the RNA polymerase, the mRNAs, and the proteins as follows:

$$\frac{d\,Pol}{d\,t} = a - \kappa \times Pol \tag{6.4a}$$

$$\frac{d\,mRNA}{d\,t} = b \times Pol - r \times mRNA \tag{6.4b}$$

$$\frac{d\,Prot}{d\,t} = c \times mRNA - \rho \times Prot \tag{6.4c}$$

These three equations each contain a production term and a degradation term for the respective pool. Following the RESET manipulation, RNA polymerases *Pol* continue to provide transcription capabilities of all mRNAs except possibly for *Pol* itself since it is specifically the β and β' genes which can be shut off. The source term a produces *Pol* from nothing but is under control of IPTG. In fact for the complex processes that produces the different subunits composing the RNA_{pol}, the approximation that this production is stoichiometric and that one does not need to consider separately all the subunits is a simplification, but nevertheless the associated modeling is probably sufficient. The gedanken experiment starts with the whole system (metabolic and coarse-grained GEM as specified by Eqs. 6.4) at its steady state, with the production terms "on". Then the IPTG is rapidly removed from the cells by dilution, which I model by setting a to 0. The RNA_{pol} is degraded at a rate κ which is obtained from the formula $\kappa = \log 2/t_{1/2}$; I choose the half life to be $t_{1/2} = 20\,min$. The mRNA source term depends on the quantity of free nucleotides, however I decided to consider a cell where the nucleotide production is indexed on the transcription requirement thus it is only the polymerase that is important to quantify the mRNA production; the corresponding rate is b. The degradation rate of the mRNAs, r, is relatively high compared to the other rates, a fact that is justified by their short half-lives of about 2 minutes. The last pool quantity is the quantity of protein (of relevance for metabolism because all enzymes are

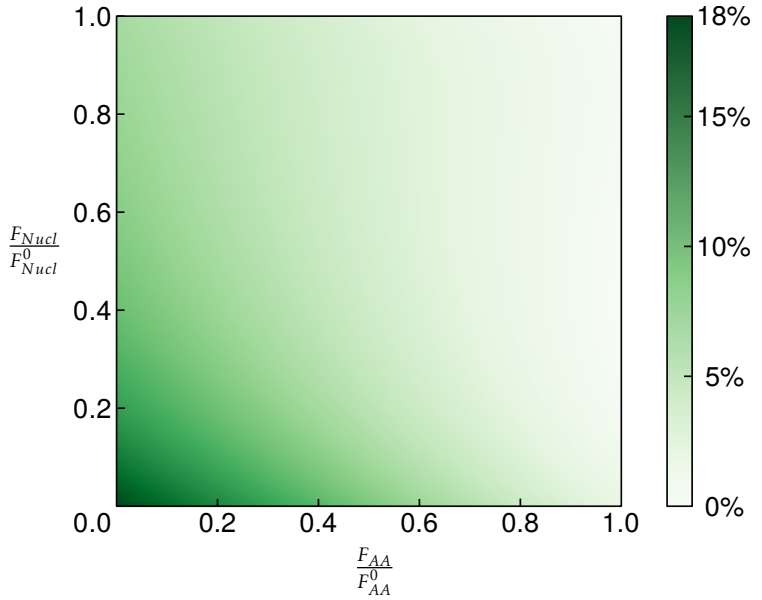


Figure 6.1: Yield in the steady state as a function of the two factors ($F_{AA}/F^0_{AA}, F_{Nucl}/F^0_{Nucl}$) relative to the yield without perturbation. These factors describe different rates of polymerization. Shown are in fact the change in yield when comparing to the reference (unperturbed) model.

proteins). For the sake of simplicity, I chose the source term for protein production to be proportional to the mRNA quantity. This assumption corresponds to neglecting changes in ribosome numbers; it is a reasonable approximation in view of the high stability of ribosomes. When the GEM is switched off, the mRNAs are produced at rates that drop with time, a decrease which feeds into the rate of translation which also drops. The protein half-life is the longest and I chose a value of $1\,h$.

To simplify the equations of Eq. 6.4, I rescale the variables to $\theta = Pol/Pol^0$, $R = mRNA/mRNA^0$, and $P = Prot/Prot^0$. The ^{0} refers to the value of the quantity in the reference state, before the perturbation is applied to the GEM. The transformed set of equations becomes

$$\frac{d\,\theta}{d\,t} = a^* - \kappa \times \theta \quad \left(a^* = \frac{a}{Pol^0}\right) \tag{6.5a}$$

$$\frac{d\,R}{d\,t} = b^* \times \theta - r \times R \quad \left(b^* = \frac{b \times Pol^0 o}{mRNA^0}\right) \tag{6.5b}$$

$$\frac{d\,P}{d\,t} = c^* \times R - \rho \times P \left(c^* = \frac{c \times mRNA^0}{Prot^0}\right) \tag{6.5c}$$

In the initial state, before any perturbation is applied, the steaty-state equations are verified and the three derivatives on the left hand side of Eq. 6.5 vanish because $\theta = R = P = 1$. From this particular state it is possible to define the values

$$\begin{aligned} a^* &= \kappa \\ b^* &= r \\ c^* &= \rho \end{aligned}$$

Based on all this, I impose the fluxes of the reactions consuming the precursors for the biomass production. In Eq. 6.3, the ratio F_{AA}/F^0_{AA} and F_{Nucl}/F^0_{Nucl} are respectively equal to R and θ. As I already mentioned, the transcription rate scales with the concentration of RNA_{pol}. Furthermore, the V_m parameters scale with the global protein abundances. The new expression for Eq. 6.3 is thus:

$$v_{out_{OAA}} = P \times ((1 - \gamma_{Nucl} - \gamma_{AA}) + \gamma_{Nucl}\theta + \gamma_{AA}R)\, V_m \frac{C_{OAA}}{K^M_{OAA} + C_{OAA}} \tag{6.6}$$

θ, R, and P evolve as described in Eqs. 6.5. After one minute, I turn on the inhibition of the gene expression machinery, so the synthesis rate of production of RNA_{pol}, a^*, is set to 0. The evolution of the yield for the model combining metabolism and the small GEM dynamical system is presented in Fig. 6.2. Interestingly the yield is not increased right away, rather it first decreases during an initial phase. One has to wait 109 min before the yield becomes larger than in the reference model. However, to become truly interesting, the modified cells must produce a larger glycerol yield when averaged over time. To search for the moment when the total glycerol yield overtakes that of the reference system, I measured the cumulative function of the two curves on Fig. 6.2. One has to wait till about 203 min for this total yield to be favorable.

The overall joint construction (kinetic model of the central model metabolism coupled to a simplified modeling of the GEM) shows important behaviours that were not predicted by the study of the steady states. The Fig. 6.1 suggests that the yield is always greater for different combinations of reductions in the amino acid and nucleotide consumption; this type of result may have been obtained from simpler stoichiometric models like an FBA framework. However, the kinetic analysis of the dynamics suggests that the yields becomes beneficial only after a not so short period of time that corresponds to a few cellular divisions for a growing bacterium. It is also important to notice that in our coarse-grained model, when the cell starts to overtake the reference glycerol production, only 15% of the enzymes remain. Although my GEM model is probably too simple to accurately describe what happens in reality, the fact that glycerol yield becomes interesing only when enzyme shortage becomes severe calls for more sophisticated models, in particular for the gene expression machinery in order to better evaluate the accuracy of these different time scales.

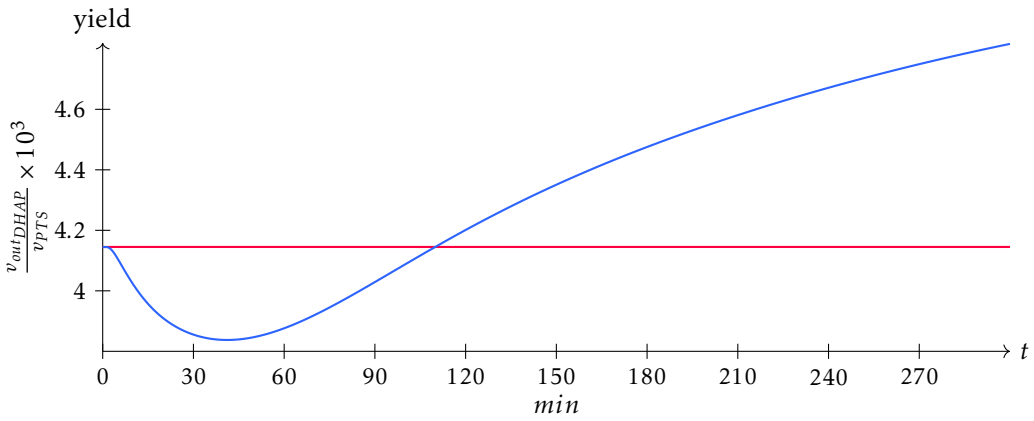


Figure 6.2: Time dependence of the glycerol yield. The model includes both metabolism in the CCM and dynamics of the gene expression machinery. The blue line represents the yield of the system when the RESET-like pertubation is applied at time 1 min. The red line shows the yield for the reference (unperturbed) model. The yield of the perturbed system overtakes the reference yield after 109 min and the RESET strategy becomes beneficial after 203 min.

7

Conclusion & Discussion

In this thesis I built a kinetic model of *E. coli* 's central carbon metabolism. The task was partly motivated by the desire to test the hypothesis of the RESET strategy, namely that by turning off the gene expression machinery, the lower demand for amino acids and nucleotides may allow the reorientation of metabolites toward other pathways. For a proof of concept, the RESET consortium is to perform such experiments on transformed *E. coli* strains and then to see the extent to which glycerol production is improved. To build my kinetic model, I had to overcome a large number of obstacles. First, an appropriate framework had to be chosen for describing the metabolism to be modeled. Because of the scientific questions posed by the RESET project, it was not possible to work with constraint-based models such as Flux Balance Analysis, instead a fully kinetic approach was necessary. Then, estimations of reaction rate parameters dispersed across the literature had to be curated. Since that search left many kinetic parameters still undetermined, I followed a Bayesian approach for adjusting the whole set of parameters in the central carbon metabolism of *E. coli* . This involved defining priors, motivated by numerous published measurements, and then exploiting systemic properties of the whole network for comparing the model's predictions to experimental measurements. At present, such a systemic approach seem the only way to build a full kinetic model since measurements on isolated individual reactions are slow coming. Finally, to implement my Bayesian framework, I had to design algorithms for obtaining in a computationally efficient way the steady state of a given kinetic model and then develop higher level algorithms for searching the parameter space of such kinetic models. The end result of these three years of work is a finalized kinetic model allowing for in silico tests of various manipulations to the metabolism in *E. coli* , in the spirit or not of the RESET project. This computational tool should be of help for metabolic engineering efforts and for providing insights into metabolic functions. A spin-off of this work was my showing how characteristic times in metabolic networks can behave in unexpected ways.

7.1 Characteristic times in metabolism

The damping of a concentration fluctuation generally requires the perturbation to spread out but in our reaction network it turns out that drift plays a central role. The time scale for evacuating a perturbation is what we call its lifetime T (cf. Eqs. 3.8 and 3.9), though in other contexts it can be referred to as the mean residence, transit or passage time. In the absence of a drift, corresponding to a pure diffusive regime, the lifetime T scales as the square of the diameter of the network, a scaling which also arises for the standard measure of return to equilibrium via the relaxation time τ. That is the situation one is most familiar with, and there τ provides the longest characteristic time as it should. However, for typical reaction networks, one has both diffusion and drift. In particular, out of equilibrium systems will have fluxes, and such fluxes may drive labeled atoms out of the network because of the associated drift. In the presence of such drift, a perturbation's lifetime T can scale as the diameter of the network divided by a characteristic drift velocity which is related to the flux intensity. Interestingly, in this out of equilibrium situation, the relaxation time τ is no longer informative about the time scale of the (slow) process which evacuates perturbations. In particular, in our toy model consisting of a homogeneous chain of reactions, τ *did not grow with the system size* while T did grow linearly. We showed analytically how that could be in that system, but the phenomenon is general. Indeed, in the presence of drift, the linearized dynamics can be decomposed into eigenvectors as without drift. The presence of drift makes the leading eigenvector (which determines τ) concentrate on the metabolites that can be excreted. As a result, τ is quite insensitive to the size of the network while T inevitably increases with network size since the evacuation of a perturbation requires it to cross the diameter of the network. These phenomena are most easily understood when the reactions obey mass action, but they arise also for Michaelis-Menten-Henri reaction laws. For this last case, the existence of a saturation of the flux with concentration of metabolites exacerbates the difference between T and τ. Interestingly, the dynamics of labeled atoms that are often used to investigate kinetic properties of networks are far less sensitive to these saturation effects. As a consequence, the use of isotopic labelings can lead one to severely underestimate the longest characteristic time in these reaction networks.

7.2 A kinetic model for central carbon metabolism

Published kinetic models appear to be quite specific to a certain choice of environmental conditions and are built with a totally different purpose than what guided me in this thesis. In the literature, the quality of a kinetic model is usually based on its ability to reproduce time courses of metabolic concentrations over short time scales (seconds to minutes). Such studies tend to downplay steady-state behavior, and in fact often do not even allow it by imposing time dependent conditions. Dropping such steady-state requirements throws away a fundamental constraint on metabolic modeling which then renders the models brittle and prevents them from being of use in other enviromental conditions. For the purposes of the RESET project, where flux reorientations are at the main issue, it is important to have both steady state behavior and reliable laws for each reaction.

My work can be considered as a first complete attempt to generate a "generic" kinetic model for the central carbon metabolism of *E. coli* , not tied to any particular experimental test. To reduce the complexity of this task, I chose to simplify a bit some of the enzymatic mechanisms, all reactions being described using the convenience kinetics formalism. This choice was made because one requires only the knowledge of the substrates and products involved in the reaction. (Of course in the future, nothing prevents one from refining this approach and inserting known allosteric regulations and the like.) The parameters in convenience kinetics are slightly different from those of detailed mechanistic formalism (cf. my chapter in which I cover the different formalisms). Another reason for not tacking yet more detailed reaction laws is that one knows that kinetic parameters can depend on conditions in the cell (pH, ionic forces) but incorporating those factors as variable would again require a much more complex model which is beyond the scope of a thesis. For all these reasons, I introduced a prior distribution for all the parameter values, using an empirical approach based on global knowledge obtained by collecting values of a same nature published in previous works. The goodness of fit of the model is then based on the agreement between the model's predictions and a number of systemic properties of *E. coli* 's central metabolism as provided in the literature. These systemic quantities are typically steady-state concentrations and fluxes, but we have added also the relaxation times of the network and the mean passage time of labelled atoms for given metabolites. The global metabolites (such as some cofactors) that intervene in many different modules of the cell metabolism are taken as parameters; I do this in particular for ATP/ADP, for CO_2, etc.

To perform the calibration (adjustment) of my kinetic model (there are several hundred parameters), I developed a genetic algorithm to search for the set of parameters that maximizes an objective function. This function is defined quantitatively for a set of parameters as the likelihood of the model given the experimental data but also given prior distributions for the parameters. The model's likelihood plays the role of its fitness in the genetic algorithm. Still, there is freedom in this objective function depending on what experimental data one wants to use or in what combinations; for instance I found that using the characteristic times and the parameter priors together with the flux and concentration distributions worked best. The prior distributions for the parameter values allow one to narrow down the parameter exploration space to "realistic" values. But these distributions are quite broad and so in practice I found they did not affect much the final adjusted values. Similarly, I found that imposing agreement with experimental characteristic times played an important role in the search algorithm, allowing it to find more quickly models with good agreement with experiments for fluxes and concentrations. The reason is that is puts a barrier to the level of saturation in the models sampled, opening better directions for searching for improved models. Indeed, if a reaction is saturated, the reaction rate is hardly affected by changes of metabolite concentrations, and so the search process has difficulty determining which way to go. Finally, this likelihood approach allows me to compute confidence intervals for each of the parameters (just as in Bayesian methods), these quantities can be measured using MCMC (Markov Chain Monte Carlo) to explore the space of models using the likelihood as the sampling measure.

The optimisation approach succeeded in providing a full kinetic model in which all the reactions and concentrations match very reasonably the systemic data we have available. The confidence intervals of the parameters are more than satisfactory, many parameters being determined within a factor 2; this was not obvious a priori given the large size of the model. Comparing to the priors, no particular surprises arose, i.e., we find hardly any outlying values for the K^M and V_m: inferred values do not deviate so

much from what is expected from the reference distributions. This may simply be due to the fact that these distributions are very broad, being spread over several orders of magnitude. Altogether, only three of these parameters were found to be outliers. Furthermore, our optimized model has reasonable relaxation times but the mean passage times were less good. That may be a consequence of our neglect of regulatory processes. Also the conditions of the experimental measurements of these times were quite different from those where the fluxes were measured.

To reduce further the uncertainties in the model's parameters, more experimental inputs would be necessary to refine our optimisation. These additional constraints could come from other types of observables such as control coefficients (the strengths of the influences of enzyme concentrations on different fluxes). Also, a knowledge of parameters of individual reactions would of course help to reduce the uncertainties of the other parameters in the model. Another word of caution follows from our taking the value of some metabolites (ATP/ADP, carbon dioxyde, etc.) as fixed. It is interesting in this regard to note that the three outlier parameters (when considering the priors) are the concentrations of phosphate, NADH and NADP. Unfortunately, given the reactions included in the CCM, there is no easy way to introduce dynamics for these concentrations. That feature is an intrinsic limitation of focusing on the central carbon metabolism which represents only a small part of the reactions involving those metabolites. An ad-hoc approach would be to let their concentration vary nevertheless using a connection to the biomass flux which consumes these metabolites.

7.3 Implications for metabolic engineering and outlooks

Once the (optimal) model constructed, I used it as an *in silico* tool to perform gedanken experiments in the spirit of the RESET strategy. The model's reactions first direct the incoming source of carbon (using the PTS for glucose influx) towards glycolysis and the pentose phosphate pathway, ultimately leading to twelve biomass compounds referred to as the essential precursors which feed into biosynthesis pathways that lead to biomass production. Among these twelve metabolites, eight are involved in the biosynthesis of amino acids and nucleotides. The gedanken experiment consists in reducing the fluxes of consumption of these essential metabolites to model the reduction in the growth rate (biomass production). That kind of change is what happens during the turning off of the gene expression machinery as implemented in the RESET project. When I do this manipulation *in silico*, I find that there is an increase in glycerol yield for the new steady state. This type of perturbation can be implemented in different ways, depending on the choice of essential precursors that are affected, leading to qualitatively similar results.

Slightly more sophisticated approaches for performing a manipulation of the gene expression machinery are possible and I investigated a minimal dynamical model to do so. I used simple ordinary differential equations describing amongst other things the decay of various species involved in the gene expression machinery such as the RNA polymerase and the mRNAs. These more detailed and realistic models also lead to an increase in glycerol yield after a certain time. This version of the gedanken experiment is a bit more appealing because it involves a pool of RNA polymerases, a pool of mRNA and a pool of proteins (the enzymes of the CCM). Each pool has a production and degradation rate. The RESET manipulation turns off the gene expression machinery; *in silico*, I implement this change by removing the source term renewing RNA polymerase. Turning off the gene expression machinery does not immediately increase the glycerol yield, it fact it is the opposite that occurs at first. During the first moments after the perturbation, the yield of glycerol decreases before ultimately increasing. The perturbed system is more efficient in terms of glycerol yield only after about 200 or so minutes. After this time lag, the cell's physiology may become critical, requiring returning to the normal situation by restarting the gene expression machinery. The simulation using my kinetic model brings out behaviours that are potentially relevant for the overall glycerol production improvement, behaviors that are not accessible if one focuses only on the steady states. It is possible that in my modeling of protein degradation my rates are overestimated; if so, the cell could spend more time in the beneficial regime which improves the yield but it is too early to confirm or infirm such a scenario. Only by quantitative estimates of the various lifetimes may one hope to answer this question.

The choice of using glycerol for a proof of concept in the RESET project emerged from the ease of

its implementation in *E. coli* . Indeed, the strain used for the RESET project has a plasmid containing the Gpd1 and Gpp2 genes; these genes from yeast code respectively for the transformation of DHAP to glycerol-3-phosphate and for the transformation of glycerol-3-phosphate to glycerol. Another, perhaps more strategic and political reason for choosing glycerol was the relatively low stakes in bioengineering for that metabolite. As a result, the project is somewhat academic, but it is its methodology that should be kept in mind. Although my *in silico* model predicts that the glycerol yeild in this system increases in the RESET restart strategy, one may note that the fluxes towards glycerol (with or without altering the gene expression machinery) are very low, only a small percentage of the carbon flux coming into the system via PTS gets channeled towards glycerol. Another choice of metabolite with higher fluxes should be used to test the effectiveness of the RESET strategy; if successful, that would provide additional credence for the usefulness of manipulating the gene expression machinery for real metabolic engineering applications.

Glossary

CF: Optimisation constraints on concentrations and on fluxes
CFP: Optimisation constraints on concentrations, on fluxes, and on parameters
CFT: Optimisation constraints on concentrations, on fluxes, and on characteristic times
CFPT: Optimisation constraints on concentrations, on fluxes, on parameters, and on characteristic times
FBA: Flux Balance Analysis
GEM: Gene expression maginery
MMH: Michaelis-Menten-Henri
TCA: Tricarboxylic Acid Cycle
PPP: Pentose phosphate Pathway

Metabolite names

BPG: 1,3-Biphospho-Glycerate
PGA2: 2-Phospho-Glycerate
PGA3: 3-Phospho-Glycerate
αKg: α-Ketoglutarate, 2-Oxoglutarate
AcCoA: Acetyl Coenzyme-A
Ace: Acetate
ADP: Adenosine Biphosphate
ATP: Adenosine Triphosphate
Cit: Citrate
CO_2: Carbon Dioxyde
CoA: Coenzyme A
DHAP: Dihydroxyacetone-Phosphate, Glycerone-Phosphate
E4P: Erythrose-4-Phosphate
F6P: Frucose-6-Phosphate
FbP: Frucose-1,6-Biphosphate
Fum: Fumarate
GAP: Glyceraldehyde-3-Phosphate
G6P: Glucose-6-Phosphate
GL6P: 6-Phosphoglucono-1,5-Lactone
Glc: Glucose
H_2O: Water
iCit: Threo-Isocitrate
KDPG: 2-Keto-3-Deoxy-6-Phospho-Gluconate, 2-Dehydro-3-Deoxy-Gluconate-6-Phosphate
Mal: Malate
NAD: Nicotinamide Adenine Dinucleotide, oxidised
NADH: Nicotinamide Adenine Dinucleotide, reduced
NADP: Nicotinamide Adenine Dinucleotide Phosphate, oxidised
NADPH: Nicotinamide Adenine Dinucleotide Phosphate, reduced
OAA: Oxaloacetate
PEP: Phosphoenolpyruvate
PGn: Gluconate-6-Phosphate
Phosph: Phosphate
Pyr: Pyruvate
R5P: Ribose-5-Phosphate
Ru5P: Ribulose-5-Phosphate
Suc: Succinate
SucCoA: Succinyl Coenzyme-A
Ubi: Ubiquinone
UbiH2: Ubiquinol

X5P: Xylulose-5-Phosphate

Reaction and enzyme names

Acn: Aconitate Hydratase
Acs: Acetate Conversion
Aldo: Aldolase
Cs: Citrate Synthase
FumA: Fumarase
PTS: Phospho Transferase System
Pgi: Phosphoglucose Isomerase
Pfk: 6-Phosphofructokinase
Eno: Enolase
Eda: 2-Keto-3-Deoxy-6-Phospho-Gluconate Aldolase
Edd: Phosphogluconate Dehydratase
Gdh: Glyceraldehyde-3-Phosphate Dehydrogenase
Gnd: 6-Phosphogluconate Dehydrogenase
Icdh: Isocitrate Deshydrogenase
Kgdh: Oxoglutarate Dehydrogenase
Mae: Malate Dehydrogenase
Mdh: Malate Oxidoreductase
out$_{\alpha Kg}$**:** αKg contribution to the Biomass
out$_{AcCoA}$**:** AcCoA contribution to the Biomass
out$_{Ace}$**:** Ace contribution to the Biomass
out$_{DHAP}$**:** DHAP contribution to the Biomass
out$_{E4P}$**:** E4P contribution to the Biomass
out$_{G6P}$**:** G6P contribution to the Biomass
out$_{R5P}$**:** R5P contribution to the Biomass
out$_{OAA}$**:** OAA contribution to the Biomass
out$_{PEP}$**:** PEP contribution to the Biomass
out$_{PGA3}$**:** PGA3 contribution to the Biomass
out$_{Pyr}$**:** Pyr contribution to the Biomass
out$_{SucCoA}$**:** SucCoA contribution to the Biomass
Pdh: Pyruvate Dehydrogenase
Pgk: Phosphoglycerate Kinase
Pgl: 6-Phosphogluconolactonase
Pgm: Phosphoglycerate Mutase
Pk: Pyruvate Kinase
Ppc: Phosphoenolpyruvate Carboxylase
Tis: Triosephosphate Isomerase
Rpe: Ribose-5-Phosphate Epimerase
Rpi: Ribose-5-Phosphate Isomerase
Sdh: Succinate Dehydrogenase
Stk: Succinate Thiokinase
Tal: Transaldolase
TktA: Transketolase I
TktB: Transketolase II
Zwf: Glucose-6-Phosphate-1-Dehydrogenase

Bibliography

[1] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, Jan 2002.

[2] Robert A Alberty. Thermodynamics in Biochemistry. *eLS*, pages 1–9, 2002.

[3] Robert A Alberty. *Biochemical thermodynamics: applications of Mathematica*, volume 48. John Wiley & Sons, 2006.

[4] Svante Arrhenius. *Über die Dissociationswärme und den Einfluss der Temperatur auf den Dissociationsgrad der Elektrolyte*. Wilhelm Engelmann, 1889.

[5] Akinori Awazu and Kunihiko Kaneko. Ubiquitous "glassy" relaxation in catalytic reaction networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 80(4):1–7, 2009.

[6] Arren Bar-even, Elad Noor, Yonatan Savir, Wolfram Liebermeister, Dan Davidi, Dan S Taw, and Ron Milo. The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters. *Biochemistry*, 50(21):4402–4410, 2011.

[7] Bryson D Bennett, Elizabeth H Kimball, Melissa Gao, Robin Osterhout, Stephen J Van Dien, and Joshua D Rabinowitz. Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. *Nature chemical biology*, 5(8):593–9, August 2009.

[8] Bryson D Bennett, Jie Yuan, Elizabeth H Kimball, and Joshua D Rabinowitz. Absolute quantitation of intracellular metabolite concentrations by an isotope ratio-based approach. *Nature protocols*, 3(8):1299–1311, 2008.

[9] Sidney W Benson and Jerry H Buss. Additivity rules for the estimation of molecular properties. thermodynamic properties. *The Journal of Chemical Physics*, 29(3):546–572, 1958.

[10] Jonathan A Bernstein, Pei-Hsun Lin, Stanley N Cohen, and Sue Lin-Chao. Global analysis of escherichia coli rna degradosome function using dna microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2758–2763, 2004.

[11] Anthony P Burgard, Priti Pharkya, and Costas D Maranas. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering*, 84(6):647–657, 2003.

[12] Ron Caspi, Tomer Altman, Richard Billington, Kate Dreher, Hartmut Foerster, Carol A Fulcher, Timothy A Holland, Ingrid M Keseler, Anamika Kothari, Aya Kubo, et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research*, 42(D1):D459–D471, 2014.

[13] C Chassagnole, D a Fell, B Raïs, B Kudla, and J P Mazat. Control of the threonine-synthesis pathway in Escherichia coli: a theoretical and experimental approach. *The Biochemical journal*, 356(Pt 2):433–444, 2001.

[14] Christophe Chassagnole, Naruemol Noisommit-Rizzi, Joachim W. Schmid, Klaus Mauch, and Matthias Reuss. Dynamic modeling of the central carbon metabolism of Escherichia coli. *Biotechnology and Bioengineering*, 79(1):53–73, July 2002.

[15] Rafael S Costa, Daniel Machado, Isabel Rocha, and Eugénio C Ferreira. Hybrid dynamic modeling of Escherichia coli central metabolic network combining Michaelis-Menten and approximate kinetic equations. *Bio Systems*, 100(2):150–7, May 2010.

[16] Markus W. Covert, Christophe H. Schilling, and Bernhard Ø Palsson. Regulation of gene expression in flux balance models of metabolism. *Journal of Theoretical Biology*, 213(1):73 – 88, 2001.

[17] Francis HC Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[18] Raul Curto, Eberhard O. Voit, Albert Sorribas, and Marta Cascante. Mathematical models of purine metabolism in man. *Mathematical Biosciences*, 151(1):1 – 49, 1998.

[19] Natalie C Duarte, Scott A Becker, Neema Jamshidi, Ines Thiele, Monica L Mo, Thuy D Vo, Rohith Srivas, and Bernhard Ø Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6):1777–1782, 2007.

[20] Natalie C Duarte, Markus J Herrgård, and Bernhard Ø Palsson. Reconstruction and validation of saccharomyces cerevisiae ind750, a fully compartmentalized genome-scale metabolic model. *Genome research*, 14(7):1298–1309, 2004.

[21] Henry Eyring. The activated complex in chemical reactions. *The Journal of Chemical Physics*, 3(2):107–115, 1935.

[22] Adam M Feist, Markus J Herrgård, Ines Thiele, Jennie L Reed, and Bernhard Ø Palsson. Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology*, 7(2):129–143, 2009.

[23] Eliane Fischer and Uwe Sauer. Metabolic flux profiling of escherichia coli mutants in central carbon metabolism using gc-ms. *European Journal of Biochemistry*, 270(5):880–891, 2003.

[24] Avi Flamholz, Elad Noor, Arren Bar-Even, Wolfram Liebermeister, and Ron Milo. Glycolytic strategy as a tradeoff between energy yield and protein cost. *Proceedings of the National Academy of Sciences of the United States of America*, 110(24):10039–44, June 2013.

[25] Hartmut Follmann and Carol Brownson. Darwin's warm little pond revisited: from molecules to the origin of life. *Naturwissenschaften*, 96(11):1265–1292, 2009.

[26] Jochen Förster, Iman Famili, Bernhard Ø Palsson, and Jens Nielsen. Large-scale evaluation of in silico gene deletions in saccharomyces cerevisiae. *OMICS A Journal of Integrative Biology*, 7(2):193–202, 2003.

[27] Hong Gao, Ye Chen, and Julie A. Leary. Kinetic measurements of phosphoglucose isomerase and phosphomannose isomerase by direct analysis of phosphorylated aldose–ketose isomers using tandem mass spectrometry. *International Journal of Mass Spectrometry*, 240(3):291 – 299, 2005. Mass Spectrometry of Biopolymers: From Model Systems to Ribosomes.

[28] Luis F. García-Alles, Alain Zahn, and Bernhard Erni. Sugar recognition by the glucose and mannose permeases of escherichia coli. steady-state kinetics and inhibition studies†. *Biochemistry*, 41(31):10077–10086, 2002. PMID: 12146972.

[29] Robert N Goldberg, Yadu B Tewari, and Talapady N Bhat. Thermodynamics of enzyme-catalyzed reactions—a database for quantitative biochemistry. *Bioinformatics*, 20(16):2874–2877, 2004.

[30] Cato M Guldberg and Peter Waage. Studies concerning affinity. *CM Forhandlinger: Videnskabs-Selskabet i Christiana*, 35, 1864.

[31] JB Haldane. *s.(1930) Enzymes*. Longmans, London, 1930.

[32] Shoshana L. Hardt. The diffusion transit time; A simple derivation. *Bulletin of Mathematical Biology*, 43:89–99, 1981.

[33] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[34] Bart R B Haverkorn van Rijsewijk, Annik Nanchen, Sophie Nallet, Roelco J Kleijn, and Uwe Sauer. Large-scale 13C-flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in Escherichia coli. *Molecular systems biology*, 7(477):477, March 2011.

[35] Victor Henri. *Lois générales de l'action des diastases*. Librairie Scientifique A. Hermann, 1903.

[36] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, , the rest of the SBML Forum:, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[37] John L Ingraham and Frederick C Neidhardt. *Escherichia coli and Salmonella: cellular and molecular biology*. ASM, 1996.

[38] François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3):318 – 356, 1961.

[39] Neema Jamshidi and Bernhard Ø Palsson. Investigating the metabolic capabilities of mycobacterium tuberculosis h37rv using the in silico strain inj661 and proposing alternative drug targets. *BMC systems biology*, 1(1):26, 2007.

[40] Neema Jamshidi and Bernhard ØPalsson. Top-Down Analysis of Temporal Hierarchy in Biochemical Reaction Networks. *PLoS computational biology*, 4(9), 2008.

[41] C.P.Leslie Grady Jr, Barth F. Smets, and Daniel S. Barbeau. Variability in kinetic parameter estimates: A review of possible causes and a proposed terminology. *Water Research*, 30(3):742 – 748, 1996.

[42] H Kacser, , and JA34 Burns. The control of flux. In *Symp. Soc. Exp. Biol.*, volume 27, pages 65–104, 1973.

[43] Tuty Asmawaty Abdul Kadir, Ahmad a Mannan, Andrzej M Kierzek, Johnjoe McFadden, and Kazuyuki Shimizu. Modeling and simulation of the main metabolism in Escherichia coli and its several single-gene knockout mutants with experimental verification. *Microbial cell factories*, 9(1):88, January 2010.

[44] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[45] T. a. Karelina, H. Ma, I. Goryanin, and O. V. Demin. EI of the Phosphotransferase System of Escherichia coli : Mathematical Modeling Approach to Analysis of Its Kinetic Properties. *Journal of Biophysics*, 2011(October 2015):1–17, 2011.

[46] Kenneth J Kauffman, Purusharth Prakash, and Jeremy S Edwards. Advances in flux balance analysis. *Current opinion in biotechnology*, 14(5):491–496, 2003.

[47] Ingrid M Keseler, Julio Collado-Vides, Alberto Santos-Zavaleta, Martin Peralta-Gil, Socorro Gama-Castro, Luis Muñiz-Rascado, César Bonavides-Martinez, Suzanne Paley, Markus Krummenacker, Tomer Altman, et al. Ecocyc: a comprehensive database of escherichia coli biology. *Nucleic acids research*, 39(suppl 1):D583–D590, 2011.

[48] Edward L. King and Carl Altman. A schematic method of deriving the rate laws for enzyme-catalyzed reactions. *The Journal of Physical Chemistry*, 60(10):1375–1378, 1956.

[49] Hans Adolf Krebs, HL Kornberg, and K Burton. *Energy transformations in living matter*. Springer, 1957.

[50] Nicolas Le Novère, Benjamin Bornstein, Alexander Broicher, Mélanie Courtot, Marco Donizelli, Harish Dharuri, Lu Li, Herbert Sauro, Maria Schilstra, Bruce Shapiro, Jacky L. Snoep, and Michael Hucka. BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Research*, 34(Database issue):D689–D691, Jan 2006.

[51] Wolfram Liebermeister and Edda Klipp. Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. *Theoretical biology & medical modelling*, 3:41, January 2006.

[52] Radhakrishnan Mahadevan, Jeremy S Edwards, and Francis J Doyle. Dynamic flux balance analysis of diauxic growth in escherichia coli. *Biophysical journal*, 83(3):1331–1340, 2002.

[53] Sebastian Meier, Pernille R. Jensen, and Jens O. Duus. Real-time detection of central carbon metabolism in living Escherichia coli and its response to perturbations. *FEBS Letters*, 585(19):3133–3138, 2011.

[54] Leonor Michaelis and Maud Menten. Die kinetik der invertinwirkung. *Biochem. z*, 49(333-369):352, 1913.

[55] Jacque Monod, Jeffries Wyman, and Jean-Pierre Changeux. On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology*, 12:88–118, 1965.

[56] Ettore Mosca, Roberta Alfieri, Carlo Maj, Annamaria Bevilacqua, Gianfranco Canti, and Luciano Milanesi. Computational modeling of the metabolic states regulated by the kinase Akt. *Frontiers in Physiology*, 3 NOV(November):1–26, 2012.

[57] Frederick C Neidhardt, John L Ingraham, and Moselio Schaechter. *Physiology of the bacterial cell: a molecular approach*. Sinauer Sunderland, 1990.

[58] Elad Noor, Hulda S Haraldsdóttir, Ron Milo, and Ronan M T Fleming. Consistent estimation of Gibbs energy using component contributions. *PLoS computational biology*, 9(7):e1003098, July 2013.

[59] L Onsager. Reciprocal Relations in Irreversible Processes. *Physical review*, 37:405–426, 1931.

[60] CFRS O'Sullivan and Frederick W Tompson. Lx.—invertase: a contribution to the history of an enzyme or unorganised ferment. *Journal of the Chemical Society, Transactions*, 57:834–931, 1890.

[61] Eleftherios T Papoutsakis. Equations and calculations for fermentations of butyric acid bacteria. *Biotechnology and bioengineering*, 26(2):174–187, 1984.

[62] Eleftherios T Papoutsakis and Charles L Meyer. Fermentation equations for propionic-acid bacteria and production of assorted oxychemicals from various sugars. *Biotechnology and bioengineering*, 27(1):67–80, 1985.

[63] Kirill Peskov, Ekaterina Mogilevskaya, and Oleg Demin. Kinetic modelling of central carbon metabolism in Escherichia coli. *The FEBS journal*, 279:3374–3385, 2012.

[64] K. Raman and N. Chandra. Flux balance analysis of biological systems: applications and challenges. *Briefings in Bioinformatics*, 10(4):435–449, 2009.

[65] P.R. Rich. The molecular machinery of keilin's respiratory chain. *Biochemical Society Transactions*, 31(6):1095–1105, 2003.

[66] B Richey, DS Cayley, MC Mossing, C Kolka, CF Anderson, TC Farrar, and MT Record. Variability of the intracellular ionic environment of escherichia coli. differences between in vitro and in vivo effects of ion concentrations on protein-dna interactions and gene expression. *Journal of Biological Chemistry*, 262(15):7157–7164, 1987.

[67] Joanne M Savinell and Bernhard O Palsson. Network analysis of intermediary metabolism using linear optimization. i. development of mathematical formalism. *Journal of theoretical biology*, 154(4):421–454, 1992.

[68] Ida Schomburg, Antje Chang, Christian Ebeling, Marion Gremse, Christian Heldt, Gregor Huhn, and Dietmar Schomburg. Brenda, the enzyme database: updates and major new developments. *Nucleic acids research*, 32(suppl 1):D431–D433, 2004.

[69] Natalie J. Stanford, Timo Lubitz, Kieran Smallbone, Edda Klipp, Pedro Mendes, and Wolfram Liebermeister. Systematic Construction of Kinetic Models from Genome-Scale Metabolic Networks. *PLoS ONE*, 8(11):e79195, 2013.

[70] Stefan Steigmiller, Paola Turina, and Peter Gräber. The thermodynamic h+/atp ratios of the h+-atpsynthases from chloroplasts and escherichia coli. *Proceedings of the National Academy of Sciences*, 105(10):3745–3750, 2008.

[71] Sattar Taheri-Araghi, Serena Bradde, John T. Sauls, Norbert S. Hill, Petra Anne Levin, Johan Paulsson, Massimo Vergassola, and Suckjoon Jun. Cell-Size Control and Homeostasis in Bacteria. *Current Biology*, 25(3):385–391, 2015.

[72] Bas Teusink, Jutta Passarge, Corinne a. Reijenga, Eugenia Esgalhado, Coen C. Van Der Weijden, Mike Schepper, Michael C. Walsh, Barbara M. Bakker, Karel Van Dam, Hans V. Westerhoff, and Jacky L. Snoep. Can yeast glycolysis be understood terms of vitro kinetics of the constituent enzymes? Testing biochemistry. *European Journal of Biochemistry*, 267(February):5313–5329, 2000.

[73] Rudolf K Thauer, Kurt Jungermann, and Karl Decker. Energy conservation in chemotrophic anaerobic bacteria. *Bacteriological reviews*, 41(1):100, 1977.

[74] M. Trauchessec, M. Jaquinod, a. Bonvalot, V. Brun, C. Bruley, D. Ropers, H. de Jong, J. Garin, G. Bestel-Corre, and M. Ferro. Mass Spectrometry-based Workflow for Accurate Quantification of Escherichia coli Enzymes: How Proteomics Can Play a Key Role in Metabolic Engineering. *Molecular & Cellular Proteomics*, 13(4):954–968, 2014.

[75] Yoshihiro Usuda, Yosuke Nishio, Shintaro Iwatani, Stephen J Van Dien, Akira Imaizumi, Kazutaka Shimbo, Naoko Kageyama, Daigo Iwahata, Hiroshi Miyano, and Kazuhiko Matsui. Dynamic modeling of Escherichia coli metabolic and regulatory systems for amino-acid production. *Journal of biotechnology*, 147(1):17–30, May 2010.

[76] Amit Varma, Brian W Boesch, and Bernhard Ø Palsson. Biochemical production capabilities of escherichia coli. *Biotechnology and bioengineering*, 42:59–73, 1993.

[77] Diana Visser and Joseph J. Heijnen. Dynamic simulation and metabolic re-design of a branched pathway using linlog kinetics. *Metabolic Engineering*, 5(3):164 – 176, 2003.

[78] Jacek Waniewski. Mean Transit Time and Mean Residence Time for Linear Diffusion–Convection–Reaction Transport System. *Computational and Mathematical Methods in Medicine*, 8(1):37–49, 2007.

[79] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[80] Jessica C Wilks and Joan L Slonczewski. ph of the cytoplasm and periplasm of escherichia coli: rapid measurement by green fluorescent protein fluorimetry. *Journal of bacteriology*, 189(15):5601–5607, 2007.

[81] Wangyun Won, Changhun Park, Sang Yup Lee, Kwang Soon Lee, and Jinwon Lee. Parameter estimation and dynamic control analysis of central carbon metabolism in Escherichia coli. *Biotechnology and Bioprocess Engineering*, 16(2):216–228, April 2011.

A

Derivation of the kinetic rate laws

A.1 Derivation of the reversible MMH equation

$$S + E \underset{k_2}{\overset{k_1}{\rightleftharpoons}} C \underset{k_4}{\overset{k_3}{\rightleftharpoons}} P + E$$

Time derivative of the reactants

$$\begin{aligned} \frac{dS}{dt} &= k_2 C - k_1 ES \\ \frac{dC}{dt} &= (k_1 ES + k_4 EP) - (k_2 + k_3)C \\ \frac{dP}{dt} &= k_3 C - k_4 EP \end{aligned} \tag{A.1}$$

Overall reaction rate

The reaction rate is given by the product variation:

$$V = \frac{dP}{dt} = k_3 C - k_4 EP \tag{A.2}$$

Assumption: the complex is at steady state

$$\frac{dC}{dt} = 0 \quad \Rightarrow \quad (k_1 ES + k_4 EP) - (k_2 + k_4)C = 0 \tag{A.3}$$

Conservation of the number of enzymes

$$E_{Tot} = E + C \tag{A.4}$$

Since we have only access to the total number of enzymes, we re-write A.3 and A.2

$$k_1 E_{Tot} S - k_1 CS + k_4 E_{Tot} P - k_4 CP - (k_2 + k_3)C = 0 \tag{A.5}$$

$$V = k_3 C - k_4 E_{Tot} P + k_4 CP \tag{A.6}$$

Derivation

Isolate C

$$\begin{aligned} C[k_1 S + k_4 P + (k_2 + k_3)] &= (k_1 S + k_4 P) E_{Tot} \\ C &= \frac{k_1 S + k_4 P}{k_1 S + k_4 P + (k_2 + k_3)} E_{Tot} \end{aligned} \tag{A.7}$$

Insertion of A.7 in A.6

$$\begin{aligned} V &= \frac{k_1 k_3 S + k_3 k_4 P}{k_1 S + k_4 P + (k_2 + k_3)} E_{Tot} - k_4 E_{Tot} P + \frac{k_1 k_4 SP + k_4^2 P^2}{k_1 S + k_4 P + (k_2 + k_3)} E_{Tot} \\ &= \left[\frac{k_1 k_3 S + k_3 k_4 P}{k_1 S + k_4 P + (k_2 + k_3)} - \frac{k_1 k_4 SP + k_4^2 P^2 + k_4 (k_2 + k_3) P}{k_1 S + k_4 P + (k_2 + k_3)} + \frac{k_1 k_4 SP + k_4^2 P^2}{k_1 S + k_4 P + (k_2 + k_3)} \right] E_{Tot} \\ &= \frac{k_1 k_3 S - k_2 k_4 P}{(k_2 + k_3) + k_1 S + k_4 P} E_{Tot} \\ &= \frac{k_1 k_3}{k_2 + k_3} \frac{S - P \frac{k_2 k_4}{k_1 k_3}}{1 + \frac{k_1}{k_2 + k_3} S + \frac{k_4}{k_2 + k_3} P} E_{Tot} \end{aligned} \tag{A.8}$$

Parameters

It is usually more convenient to redefine the parameters in order to look more like a Michaelis-Menten equation.

Saturation constants or Michaelis constants

$$K_m^S = \frac{k_2 + k_3}{k_1} \qquad K_m^P = \frac{k_2 + k_3}{k_4}$$

Forward and backward reaction rates

$$k^+ = k_3 \qquad k^- = k_2$$

Equilibrium constant

$$K_{eq} = \frac{k_2 k_4}{k_1 k_3} = \frac{k^+ K_m^P}{k^- K_m^S}$$

Final equation

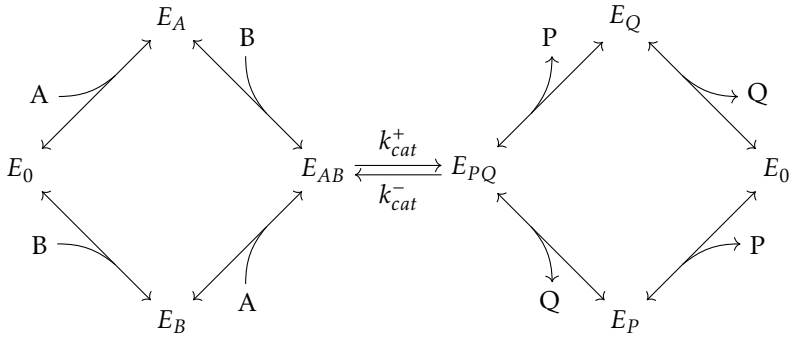
$$\boxed{V = k^+/K_m^S \times \frac{S - P/K_{eq}}{1 + S/K_m^S + P/K_m^P} E_{Tot}}$$

A.2 Derivation of the convenience kinetic rate law

To derive convenience rate laws, several assumption need to be made:

1. The substrates and products can respectively bind and unbind to an enzyme in any order. In other word, only random mechanism are considered.
2. Binding of substrates and unbinding of products is much faster than the conversion of the central complex, this assumption is the same as considering the binding and unbinding steps at equilibrium.
3. The binding energy energy of individual reactant does not depend on the precence or not of enzymes on the enzyme.

As a working example let us consider the case where two substrate and two products interact:



Notation $[E_x]$ stands for the concentration of the complex E_x and x denotes the concentration of the metabolite X.

Equilibrium of the binding and unbinding reactions

Equilibrium of

$$A + E_0 \rightleftarrows E_A$$

Which results in the Haldane equation

$$\frac{[E_A]}{[E_0]} = \frac{a}{k_M^A} = \tilde{a}$$

where k_M^A is the dissociation constant by definition. The same manipulation is done to define $\tilde{b}$,$\tilde{p}$, and $\tilde{y}$.

And easily one finds:

$$\frac{E_{AB}}{E_0} = \tilde{a} \times \tilde{b} \qquad \frac{E_{PQ}}{E_0} = \tilde{p} \times \tilde{q} \tag{A.9}$$

The conversion reaction imposes the reaction's rate

Be cause the other reaction are in fast equilibrium, the conversion reaction imposes the reaction's rate:

$$r = k_{cat}^{+}[E_{AB}] - k_{cat}^{-}[E_{PQ}] = (k_{cat}^{+}\tilde{a}\tilde{b} - k_{cat}^{-}p\tilde{q})[E_0] \tag{A.10}$$

Introduction of the total enzyme concentration

It is always more convenient to deal with the total enzyme concentration that is known than only the free enzymes. To do so, it is necessary to replace define the total concentration as the sum of all the enzyme complexes plus the free enzyme:

$$[E_{Tot}] = [E_0](1 + \tilde{a} + \tilde{b} + \tilde{a}\tilde{b} + \tilde{p} + \tilde{q} + \tilde{p}\tilde{q})$$
$$r = \frac{k_{cat}^{+}\tilde{a}\tilde{b} - k_{cat}^{-}\tilde{p}\tilde{q}}{(1+\tilde{a})(1+\tilde{b}) + (1+\tilde{p})(1+\tilde{q}) - 1}[E_{Tot}]$$

A similar equation can be derived in a general case:

$$r = \frac{k_{cat}^{+}\prod_i \tilde{s}_i - k_{cat}^{-}\prod_j \tilde{p}_j}{\prod_i(1+\tilde{s}_i) + \prod_j(1+\tilde{p}_j) - 1}[E_{Tot}] \tag{A.11}$$

Action of inhibitors

To take into account the action of inhibitors, those are considered to act via mixed inhibition, thus to represent their influence one can multiply the whole rate by a factor

$$\frac{k_I}{i + k_I} \tag{A.12}$$

where i represents the inhibitor's concentration and k_I the inhibition constant.

Action of activators

Similarly, the action of molecules that enhance the effficacity of an enzyme can be modeled by multiplying the equation A.11 by

$$\frac{a}{k_A + a}$$

where a in the activator's concentration and k_A the activation constant.

B

Reference concentrations and fluxes

Concentrations

Metabolite	Concentrion	min 95%	max 95%
PEP	0.184	0.146	0.231
G6P + F6P	8.75	1.57	15.1
ATP	9.63	8.13	11.4
ADP	0.555	0.437	0.644
FbP	15.2	14	16.4
GAP + DHAP	0.374	0.344	0.405
NAD	2.55	2.32	2.80
NADH	0.0832	0.0545	0.127
PGA3 + PGA2	1.54	1.51	1.58
NADP	0.00208	0.000147	0.0311
GL6P	1.04	0.647	1.68
NADPH	0.121	0.11	0.134
PGn	3.77	3.67	3.85
Ru5P +R5P+X5P	1.32	0.983	1.77
AcCoA	0.606	0.529	0.694
Icit + Cit	1.96	1.10	3.48
aKg	0.443	0.312	0.631
SucCoA	0.23	0.142	0.383
Suc	0.569	0.341	0.949
Fum	0.115	0.03	4.42
Mal	1.68	1.66	1.7

Table B.1: Concentrations from Bennett & al [7]. The concentration values and the 95 % intervals are in mM.

B.1 Fluxes

Flux	Value	95%interval
PTS	8.13	0.17
Pgi	5.66	0.195
Pfk	6.46	0.31
Aldo	6.46	0.31
Tis	6.4	0.31
Gdh	13.87	0.42
Pgk	13.87	0.42
Pgm	12.94	0.42
Eno	12.94	0.42
Pk	0.86	0.49
Zwf	2.39	0.195
Pgl	2.39	0.195
Gnd	1.65	0.22
Rpi	0.85	0.37
Rpe	0.8	0.15
TktA	0.27	0.075
TktB	0.53	0.075
Tal	0.53	0.075
Edd	0.74	0.324
Eda	0.74	0.325
Pdh	9.14	0.64
Acs	5.48	0.57
Ppc	2.40	1.04
Cs	2.20	0.45
Acn	2.20	0.45
Icdh	2.20	0.45
Stk	1.29	0.44
Sdh	1.29	0.44
FumA	1.29	0.44
Mdh	0.81	0.5
Mae	0.48	0.73
outG6P	0.08	0.56
outDHAP	0.06	0.62
outR5P	0.32	0.445
outE4P	0.26	0.15
outPGA3	0.93	0.84
outPEP	0.54	2.12
outPyr	2.07	1.598
outAcCoA	1.46	1.66
outAce	5.48	0.57
outOAA	1.02	1.99

Table B.2: Reference fluxes from Haverkorn van Rijsewijk & al [34]. The fluxes and the 95% intervals are in $mmol\ gDCW^{-1}$.

C

Optimized parameters and confidence intervals

This annex gathers the values of the parameters and their coefficient of variations for the three the four optimisation CF, CFP, CFT, and CFPT. The number into bracket are the coefficient of variation. The units are mM for the concentrations and the K^Ms, $mM\,l^{-1}$ for the rate maximum velocities, J for the energies of formations.

Parameter	CF	CFP	CFT	CFPT
PTS: V_m	10.219(0.150)	2.269(0.220)	18.445(0.146)	2.405(0.268)
PTS: K^M_{Glc}	0.270(0.242)	0.136(0.151)	0.403(0.293)	0.145(0.144)
PTS: K^M_{PEP}	0.318(0.227)	0.114(0.156)	0.342(0.230)	0.113(0.155)
PTS: K^M_{G6P}	0.190(0.142)	0.265(0.229)	0.123(0.170)	0.244(0.275)
PTS: K^M_{Pyr}	0.136(0.171)	0.265(0.265)	0.160(0.172)	0.261(0.197)
Pgi: V_m	1.139(0.812)	1.376(0.527)	1.154(0.679)	1.381(0.726)
Pgi: K^M_{G6P}	0.578(0.751)	0.136(0.169)	0.683(0.680)	0.137(0.159)
Pgi: K^M_{F6P}	0.175(0.268)	0.257(0.260)	0.133(0.190)	0.256(0.248)
Pgi: K^i_{PEP}	0.335(0.305)	0.223(0.188)	0.358(0.275)	0.228(0.272)
Pfk: V_m	14.726(0.222)	1.995(0.487)	429.358(0.076)	2.274(0.472)
Pfk: K^M_{F6P}	0.088(0.141)	0.098(0.124)	0.009(0.067)	0.098(0.098)
Pfk: K^M_{ATP}	0.118(0.134)	0.116(0.142)	0.098(0.168)	0.115(0.136)
Pfk: K^M_{ADP}	0.210(0.291)	0.301(0.307)	0.343(0.269)	0.308(0.384)
Pfk: K^M_{FbP}	0.294(0.362)	0.330(0.251)	0.318(0.300)	0.333(0.360)
Aldo: V_m	0.575(0.371)	0.557(0.346)	0.493(0.279)	0.563(0.332)
Aldo: K^M_{FbP}	0.401(0.323)	0.146(0.218)	0.366(0.241)	0.147(0.169)
Aldo: K^M_{DHAP}	0.136(0.148)	0.247(0.321)	0.406(0.355)	0.246(0.264)
Aldo: K^M_{GAP}	0.116(0.244)	0.264(0.219)	0.240(0.302)	0.258(0.251)
Aldo: K^a_{aKg}	0.223(0.162)	0.183(0.153)	0.299(0.230)	0.183(0.210)
Tis: V_m	441.780(0.074)	5.316(0.351)	16898.108(0.049)	5.284(0.379)
Tis: K^M_{DHAP}	1.310(0.894)	0.112(0.091)	1.228(1.320)	0.111(0.101)
Tis: K^M_{GAP}	0.248(0.226)	0.271(0.249)	0.287(0.215)	0.271(0.262)
Gdh: V_m	34783.219(0.056)	15.705(0.133)	728619.000(0.050)	16.065(0.139)
Gdh: K^M_{GAP}	0.879(3.504)	0.061(0.123)	0.900(1.943)	0.061(0.116)
Gdh: K^M_{Phosph}	0.550(0.375)	0.141(0.108)	0.492(0.299)	0.137(0.150)
Gdh: K^M_{NAD}	0.024(0.102)	0.133(0.127)	0.019(0.062)	0.130(0.148)
Gdh: K^M_{BPG}	0.231(0.312)	0.258(0.334)	0.322(0.265)	0.258(0.273)
Gdh: K^M_{NADH}	0.260(0.256)	0.255(0.307)	0.296(0.243)	0.255(0.241)
Pgk: V_m	158.350(0.062)	12.338(0.179)	710.443(0.070)	13.138(0.187)
Pgk: K^M_{BPG}	0.194(0.152)	0.052(0.123)	0.240(0.248)	0.051(0.131)
Pgk: K^M_{ADP}	0.199(0.247)	0.062(0.102)	0.239(0.223)	0.061(0.120)
Pgk: K^M_{PGA3}	0.097(0.140)	0.587(0.583)	0.159(0.138)	0.601(0.700)
Pgk: K^M_{ATP}	0.209(0.157)	0.743(2.407)	0.116(0.139)	0.722(0.619)
Pgm: V_m	1594.136(0.061)	3.529(0.189)	2448.831(0.054)	3.588(0.277)
Pgm: K^M_{PGA3}	0.137(0.239)	0.132(0.142)	0.171(0.211)	0.127(0.182)
Pgm: K^M_{PGA2}	0.318(0.336)	0.267(0.250)	0.166(0.151)	0.274(0.198)
Eno: V_m	4.825(0.156)	4.486(0.260)	4.923(0.167)	4.503(0.208)
Eno: K^M_{PGA2}	0.217(0.197)	0.089(0.137)	0.257(0.301)	0.091(0.132)
Eno: K^M_{PEP}	0.200(0.294)	0.321(0.298)	0.283(0.290)	0.320(0.221)
Pk: V_m	3.475(0.388)	1.376(0.855)	3.771(0.391)	1.238(1.203)
Pk: K^M_{PEP}	0.424(0.283)	0.090(0.102)	0.346(0.375)	0.092(0.141)
Pk: K^M_{ADP}	0.228(0.243)	0.094(0.091)	0.302(0.257)	0.102(0.153)
Pk: K^M_{Pyr}	0.173(0.268)	0.338(0.431)	0.240(0.228)	0.339(0.338)
Pk: K^M_{ATP}	12.793(0.154)	0.718(1.167)	8.733(0.214)	0.815(1.273)
Pk: K^a_{FbP}	0.248(0.237)	0.195(0.308)	0.209(0.206)	0.184(0.148)
Zwf: V_m	2.441(0.411)	2.874(0.445)	2.567(0.341)	3.174(0.216)

Zwf: K^M_{G6P}	0.159(0.184)	0.135(0.163)	0.160(0.137)	0.135(0.169)
Zwf: K^M_{NADP}	0.013(0.074)	0.082(0.084)	0.012(0.075)	0.080(0.119)
Zwf: K^M_{GL6P}	0.245(0.264)	0.277(0.286)	0.265(0.185)	0.282(0.303)
Zwf: K^M_{NADPH}	0.202(0.300)	0.260(0.210)	0.287(0.375)	0.261(0.362)
Pgl: V_m	1.313(0.883)	0.793(1.983)	1.228(1.447)	0.875(5.541)
Pgl: K^M_{GL6P}	0.182(0.223)	0.132(0.137)	0.224(0.182)	0.131(0.159)
Pgl: K^M_{PGn}	0.192(0.157)	0.285(0.423)	0.298(0.253)	0.278(0.306)
Gnd: V_m	9.775(0.165)	2.086(0.477)	12.206(0.087)	2.166(0.415)
Gnd: K^M_{PGn}	0.107(0.175)	0.128(0.164)	0.063(0.133)	0.128(0.120)
Gnd: K^M_{NADP}	0.109(0.118)	0.085(0.096)	0.090(0.091)	0.081(0.090)
Gnd: K^M_{Ru5P}	0.320(0.401)	0.278(0.245)	0.319(0.221)	0.287(0.194)
Gnd: K^M_{NADPH}	0.194(0.200)	0.264(0.296)	0.330(0.257)	0.263(0.306)
Gnd: K^M_{CO2}	0.332(0.282)	0.277(0.283)	0.349(0.306)	0.270(0.198)
Rpi: V_m	35.830(0.198)	0.152(0.184)	2.160(0.717)	0.565(1.342)
Rpi: K^M_{Ru5P}	3.329(0.271)	0.165(0.147)	7.892(0.131)	0.150(0.202)
Rpi: K^M_{R5P}	2.383(0.346)	0.252(0.259)	1.925(0.420)	0.493(0.639)
Rpe: V_m	5.470(0.340)	0.747(17.068)	1.480(0.515)	0.882(1.268)
Rpe: K^M_{Ru5P}	1.803(0.397)	0.136(0.121)	2.374(0.403)	0.133(0.150)
Rpe: K^M_{X5P}	0.178(0.150)	0.277(0.313)	0.019(0.094)	0.269(0.280)
TktA: V_m	0.064(0.132)	0.118(0.258)	0.094(0.198)	0.108(0.224)
TktA: K^M_{X5P}	0.163(0.164)	0.149(0.148)	0.138(0.142)	0.152(0.192)
TktA: K^M_{E4P}	0.129(0.192)	0.180(0.206)	0.161(0.120)	0.181(0.140)
TktA: K^M_{GAP}	21.793(0.131)	0.264(0.271)	36.082(0.084)	0.472(0.585)
TktA: K^M_{F6P}	1.118(9.629)	0.256(0.339)	1.101(1.941)	0.483(0.671)
TktB: V_m	0.272(0.325)	0.361(0.791)	0.455(0.509)	0.105(0.170)
TktB: K^M_{R5P}	1.420(0.954)	0.135(0.181)	1.344(1.112)	0.180(0.145)
TktB: K^M_{X5P}	0.150(0.176)	0.139(0.118)	0.093(0.096)	0.162(0.180)
TktB: K^M_{GAP}	10.258(0.152)	0.269(0.298)	19.091(0.097)	0.501(0.533)
TktB: K^M_{S7P}	4.186(0.441)	0.282(0.313)	7.004(0.141)	0.514(0.771)
Tal: V_m	9.166(0.241)	0.797(5.094)	6.984(0.193)	0.857(3.838)
Tal: K^M_{S7P}	0.313(0.275)	0.135(0.203)	0.294(0.171)	0.146(0.142)
Tal: K^M_{GAP}	0.959(7.054)	0.117(0.108)	0.947(1.930)	0.151(0.176)
Tal: K^M_{E4P}	0.217(0.317)	0.262(0.226)	0.106(0.164)	0.247(0.240)
Tal: K^M_{F6P}	1.525(1.213)	0.283(0.275)	1.443(1.177)	0.527(0.575)
Pdh: V_m	7.245(0.283)	2.582(0.250)	10.522(0.148)	2.712(0.248)
Pdh: K^M_{Pyr}	0.419(0.470)	0.119(0.155)	0.230(0.189)	0.119(0.194)
Pdh: K^M_{CoA}	0.022(0.089)	0.113(0.187)	0.031(0.050)	0.112(0.134)
Pdh: K^M_{NAD}	0.118(0.140)	0.134(0.185)	0.087(0.095)	0.138(0.145)
Pdh: K^M_{CO2}	0.224(0.197)	0.253(0.273)	0.275(0.234)	0.260(0.249)
Pdh: K^M_{AcCoA}	0.368(0.300)	0.254(0.186)	0.340(0.339)	0.261(0.325)
Pdh: K^M_{NADH}	0.201(0.276)	0.255(0.221)	0.220(0.252)	0.259(0.203)
Pta: V_m	7.773(0.198)	1.168(1.132)	7.433(0.208)	1.202(0.969)
Pta: K^M_{Phosph}	1.924(0.584)	0.139(0.151)	2.007(0.539)	0.140(0.154)
Pta: K^M_{ADP}	0.156(0.189)	0.131(0.203)	0.153(0.128)	0.142(0.115)
Pta: K^M_{AcCoA}	0.052(0.107)	0.135(0.166)	0.031(0.080)	0.138(0.155)
Pta: K^M_{ATP}	0.231(0.362)	0.261(0.233)	0.248(0.270)	0.255(0.310)
Pta: K^M_{CoA}	0.311(0.270)	0.257(0.224)	0.188(0.275)	0.271(0.324)
Pta: K^M_{Ace}	0.196(0.142)	0.256(0.262)	0.198(0.252)	0.262(0.328)
Ppc: V_m	30.616(0.139)	4.635(0.320)	30.728(0.112)	4.393(0.327)
Ppc: K^M_{PEP}	0.453(0.384)	0.071(0.095)	0.467(0.281)	0.070(0.103)
Ppc: K^M_{CO2}	0.137(0.189)	0.074(0.111)	0.094(0.113)	0.076(0.104)

Ppc: K^M_{Phosph}	0.322(0.252)	0.562(0.561)	0.207(0.255)	0.515(0.550)
Ppc: K^M_{OAA}	0.280(0.290)	0.269(0.178)	0.301(0.266)	0.269(0.320)
Ppc: K^i_{Mal}	0.176(0.200)	0.434(0.631)	0.168(0.171)	0.445(0.397)
Ppc: K^a_{AcCoA}	0.178(0.299)	0.251(0.232)	0.160(0.203)	0.253(0.188)
Ppc: K^i_{FbP}	0.148(0.239)	0.189(0.218)	0.290(0.231)	0.190(0.221)
Cs: V_m	1.884(0.489)	3.509(0.308)	3.704(0.336)	3.140(0.269)
Cs: K^M_{OAA}	0.225(0.159)	0.067(0.101)	0.171(0.186)	0.065(0.106)
Cs: K^M_{AcCoA}	0.285(0.222)	0.092(0.087)	0.335(0.305)	0.091(0.148)
Cs: K^M_{CoA}	0.140(0.191)	0.318(0.332)	0.130(0.194)	0.308(0.226)
Cs: K^M_{Cit}	0.143(0.211)	0.410(0.300)	0.085(0.115)	0.405(0.230)
Acn: V_m	71.492(0.148)	1.173(2.462)	146.446(0.108)	1.223(1.403)
Acn: K^M_{Cit}	8.522(0.190)	0.137(0.140)	4.197(0.179)	0.139(0.149)
Acn: K^M_{Icit}	27.239(0.105)	0.279(0.185)	36.565(0.067)	0.557(0.659)
Icdh: V_m	1044.345(0.069)	9.902(0.161)	700424.580(0.036)	9.663(0.167)
Icdh: K^M_{Icit}	0.037(0.118)	0.061(0.088)	0.019(0.093)	0.057(0.077)
Icdh: K^M_{NADP}	0.032(0.075)	0.049(0.099)	0.037(0.075)	0.048(0.105)
Icdh: K^M_{NADPH}	0.168(0.203)	0.337(0.275)	0.226(0.352)	0.333(0.276)
Icdh: K^M_{CO2}	0.295(0.301)	0.445(0.436)	0.137(0.186)	0.426(0.428)
Icdh: K^M_{aKg}	0.202(0.252)	0.431(0.424)	0.262(0.203)	0.421(0.488)
Kgdh: V_m	0.790(1.739)	0.555(0.483)	0.835(6.720)	0.681(0.692)
Kgdh: K^M_{aKg}	0.023(0.076)	0.151(0.161)	0.009(0.080)	0.166(0.259)
Kgdh: K^M_{CoA}	0.026(0.076)	0.153(0.119)	0.025(0.069)	0.171(0.185)
Kgdh: K^M_{NAD}	0.089(0.133)	0.142(0.169)	0.078(0.097)	0.149(0.131)
Kgdh: K^M_{CO2}	0.336(0.553)	0.257(0.293)	0.245(0.161)	0.249(0.335)
Kgdh: K^M_{SucCoA}	0.283(0.361)	0.256(0.227)	0.190(0.170)	0.256(0.273)
Kgdh: K^M_{NADH}	0.202(0.294)	0.262(0.278)	0.322(0.396)	0.262(0.248)
Kgdh: K^i_{OAA}	0.286(0.290)	0.190(0.226)	0.322(0.209)	0.185(0.171)
Stk: V_m	800.767(0.110)	0.915(7.945)	721.917(0.067)	0.957(2.126)
Stk: K^M_{SucCoA}	0.106(0.168)	0.131(0.111)	0.158(0.135)	0.132(0.152)
Stk: K^M_{Phosph}	0.183(0.199)	0.140(0.145)	0.132(0.128)	0.141(0.188)
Stk: K^M_{ADP}	0.132(0.184)	0.136(0.142)	0.218(0.145)	0.139(0.145)
Stk: K^M_{CoA}	0.058(0.097)	0.265(0.192)	0.092(0.115)	0.264(0.258)
Stk: K^M_{Suc}	0.289(0.366)	0.259(0.247)	0.261(0.174)	0.252(0.310)
Stk: K^M_{ATP}	0.085(0.116)	0.265(0.216)	0.093(0.171)	0.250(0.375)
Sdh: V_m	2.970(0.449)	0.264(0.285)	10.761(0.165)	0.275(0.564)
Sdh: K^M_{Suc}	0.013(0.087)	0.147(0.137)	0.008(0.059)	0.147(0.195)
Sdh: K^M_{Ubi}	0.144(0.124)	0.147(0.202)	0.135(0.132)	0.152(0.200)
Sdh: K^M_{Fum}	0.000(0.039)	0.244(0.221)	0.000(0.037)	0.246(0.274)
Sdh: K^M_{UbiH2}	0.218(0.247)	0.242(0.259)	0.280(0.174)	0.242(0.263)
FumA: V_m	47813.708(0.055)	3.947(0.247)	36350.493(0.035)	3.981(0.418)
FumA: K^M_{Fum}	0.185(0.160)	0.092(0.149)	0.243(0.182)	0.096(0.144)
FumA: K^M_{Mal}	0.656(0.533)	0.384(0.330)	0.640(0.743)	0.400(0.382)
FumA: K^i_{Cit}	0.436(0.415)	0.411(0.379)	0.384(0.542)	0.410(0.325)
Mdh: V_m	1.571(1.048)	1.064(3.900)	1.898(0.562)	1.142(1.761)
Mdh: K^M_{NAD}	0.254(0.228)	0.140(0.214)	0.383(0.265)	0.139(0.155)
Mdh: K^M_{Mal}	3.831(0.226)	0.140(0.166)	2.541(0.276)	0.139(0.123)
Mdh: K^M_{OAA}	0.008(0.096)	0.253(0.215)	0.003(0.049)	0.245(0.317)
Mdh: K^M_{NADH}	0.066(0.136)	0.267(0.217)	0.072(0.135)	0.249(0.256)
Mae: V_m	2.388(0.774)	0.415(0.482)	2.623(0.426)	0.369(0.601)
Mae: K^M_{Mal}	0.195(0.260)	0.140(0.135)	0.200(0.163)	0.137(0.132)
Mae: K^M_{NAD}	0.240(0.286)	0.142(0.201)	0.149(0.149)	0.143(0.203)

Mae: K^M_{CO2}	0.189(0.272)	0.255(0.310)	0.170(0.231)	0.260(0.233)
Mae: K^M_{Pyr}	0.121(0.188)	0.257(0.252)	0.225(0.273)	0.249(0.265)
Mae: K^M_{NADH}	0.314(0.355)	0.259(0.280)	0.223(0.172)	0.255(0.262)
Mae: K^a_{G6P}	0.247(0.303)	0.187(0.157)	0.225(0.200)	0.183(0.217)
Mae: K^i_{Fum}	0.095(0.169)	0.192(0.206)	0.085(0.082)	0.181(0.268)
Mae: K^i_{OAA}	0.165(0.215)	0.188(0.204)	0.182(0.237)	0.186(0.139)
Mae: K^i_{AcCoA}	0.155(0.178)	0.195(0.240)	0.075(0.118)	0.184(0.184)
Edd: V_m	0.096(0.182)	0.099(0.277)	0.104(0.150)	0.104(0.161)
Edd: K^M_{PGn}	0.145(0.161)	0.147(0.139)	0.246(0.148)	0.146(0.158)
Edd: K^M_{KDPG}	0.224(0.264)	0.241(0.325)	0.227(0.236)	0.253(0.237)
Eda: V_m	0.119(0.179)	0.202(0.233)	0.143(0.231)	0.226(0.319)
Eda: K^M_{KDPG}	0.111(0.111)	0.158(0.167)	0.115(0.172)	0.150(0.150)
Eda: K^M_{GAP}	0.199(0.214)	0.250(0.290)	0.255(0.274)	0.253(0.234)
Eda: K^M_{Pyr}	9.912(0.107)	0.248(0.347)	8.095(0.170)	0.515(0.772)
out_{G6P}: V_m	0.001(0.058)	0.007(0.098)	0.001(0.078)	0.007(0.099)
out_{G6P}: K^M_{G6P}	0.111(0.137)	0.148(0.177)	0.142(0.172)	0.146(0.144)
out_{DHAP}: V_m	0.000(0.069)	0.007(0.076)	0.000(0.096)	0.007(0.107)
out_{DHAP}: K^M_{DHAP}	0.123(0.130)	0.240(0.208)	0.145(0.110)	0.235(0.217)
out_{R5P}: V_m	0.042(0.164)	0.163(0.296)	0.098(0.208)	0.102(0.211)
out_{R5P}: K^M_{R5P}	0.206(0.192)	0.173(0.197)	0.187(0.165)	0.167(0.199)
out_{E4P}: V_m	0.024(0.164)	0.074(0.183)	0.029(0.105)	0.096(0.218)
out_{E4P}: K^M_{E4P}	0.155(0.157)	0.199(0.207)	0.110(0.080)	0.201(0.152)
out_{PGA3}: V_m	0.219(0.316)	0.146(0.284)	0.219(0.237)	0.156(0.227)
out_{PGA3}: K^M_{PGA3}	0.987(12.957)	0.142(0.254)	0.997(5.174)	0.150(0.125)
out_{PEP}: V_m	0.136(0.218)	0.105(0.340)	0.133(0.277)	0.090(0.218)
out_{PEP}: K^M_{PEP}	0.959(3.272)	0.178(0.178)	0.985(2.040)	0.174(0.159)
out_{Pyr}: V_m	0.329(0.345)	0.357(0.394)	0.340(0.347)	0.370(0.442)
out_{Pyr}: K^M_{Pyr}	0.101(0.153)	0.139(0.198)	0.121(0.137)	0.142(0.137)
out_{AcCoA}: V_m	0.185(0.273)	0.201(0.369)	0.189(0.181)	0.198(0.289)
out_{AcCoA}: K^M_{AcCoA}	0.160(0.229)	0.147(0.170)	0.161(0.105)	0.145(0.154)
out_{Ace}: V_m	1.122(1.410)	0.920(7.767)	1.143(2.373)	0.940(3.761)
out_{Ace}: K^M_{Ace}	0.235(0.177)	0.134(0.112)	0.236(0.193)	0.136(0.167)
out_{OAA}: V_m	0.393(0.533)	0.755(1.763)	0.811(0.885)	0.812(2.513)
out_{OAA}: K^M_{OAA}	0.968(3.034)	0.124(0.172)	0.974(2.435)	0.133(0.136)
out_{aKg}: V_m	0.237(0.192)	0.134(0.228)	0.195(0.201)	0.124(0.202)
out_{aKg}: K^M_{aKg}	0.093(0.146)	0.160(0.204)	0.043(0.095)	0.156(0.163)
out_{SucCoA}: V_m	0.005(0.104)	0.176(0.269)	0.001(0.086)	0.189(0.257)
out_{SucCoA}: K^M_{SucCoA}	0.140(0.200)	0.159(0.163)	0.179(0.145)	0.156(0.236)
Glc	16.700(0.000)	16.700(0.000)	16.700(0.000)	16.700(0.000)
ATP	8.272(0.021)	9.363(0.015)	13.366(0.014)	9.340(0.016)
ADP	0.574(0.123)	0.571(0.114)	0.507(0.107)	0.580(0.108)
Phosph	0.919(55.767)	18.442(0.086)	0.984(59.595)	15.711(0.143)
NAD	2.605(0.026)	2.599(0.021)	2.658(0.024)	2.599(0.022)
NADH	0.000(0.016)	0.006(0.043)	0.000(0.023)	0.006(0.035)
NADP	0.003(0.049)	0.010(0.087)	0.002(0.046)	0.009(0.061)
NADPH	0.104(0.061)	0.096(0.045)	0.100(0.046)	0.096(0.043)
CO2	0.495(0.689)	0.367(0.448)	0.486(0.459)	0.366(0.386)
CoA	0.012(0.069)	0.188(0.194)	0.007(0.071)	0.162(0.160)
Ubi	0.339(0.401)	0.390(0.521)	0.294(0.287)	0.384(0.497)
UbiH2	0.319(0.291)	0.446(0.405)	0.337(0.316)	0.494(0.388)
$\Delta_f G_{Glc}$	-410199.366(0.000)	-410199.997(0.000)	-410199.818(0.000)	-410200.171(0.000)
$\Delta_f G_{PEP}$	-1203100.886(0.000)	-1203099.975(0.000)	-1203100.532(0.000)	-1203100.041(0.000)

$\Delta_f G_{G6P}$	-1303400.441(0.000)	-1303400.779(0.000)	-1303399.641(0.000)	-1303401.030(0.000)
$\Delta_f G_{Pyr}$	-354000.185(0.000)	-354000.577(0.000)	-354000.366(0.000)	-354000.504(0.000)
$\Delta_f G_{F6P}$	-1300900.677(0.000)	-1300900.134(0.000)	-1300899.155(0.000)	-1300899.969(0.000)
$\Delta_f G_{ATP}$	-2278899.951(0.000)	-2278902.019(0.000)	-2278902.216(0.000)	-2278902.247(0.000)
$\Delta_f G_{ADP}$	-2193600.337(0.000)	-2193599.513(0.000)	-2193598.130(0.000)	-2193599.311(0.000)
$\Delta_f G_{FbP}$	-1405001.823(0.000)	-1405000.531(0.000)	-1405000.665(0.000)	-1405000.563(0.000)
$\Delta_f G_{DHAP}$	-1097300.200(0.000)	-1097299.200(0.000)	-1097299.207(0.000)	-1097299.282(0.000)
$\Delta_f G_{GAP}$	-1091800.670(0.000)	-1091799.800(0.000)	-1091801.171(0.000)	-1091799.943(0.000)
$\Delta_f G_{Phosph}$	-1070599.446(0.000)	-1070598.711(0.000)	-1070600.205(0.000)	-1070598.708(0.000)
$\Delta_f G_{NAD}$	-1142299.388(0.000)	-1142298.278(0.000)	-1142299.942(0.000)	-1142298.457(0.000)
$\Delta_f G_{BPG}$	-2209799.339(0.000)	-2209803.282(0.000)	-2209799.358(0.000)	-2209803.399(0.000)
$\Delta_f G_{NADH}$	-1075700.929(0.000)	-1075702.057(0.000)	-1075700.583(0.000)	-1075701.884(0.000)
$\Delta_f G_{PGA3}$	-1354700.583(0.000)	-1354699.462(0.000)	-1354699.582(0.000)	-1354700.006(0.000)
$\Delta_f G_{PGA2}$	-1350499.495(0.000)	-1350501.886(0.000)	-1350498.980(0.000)	-1350501.654(0.000)
$\Delta_f G_{H2O}$	-151499.952(0.000)	-151500.153(0.000)	-151499.898(0.000)	-151500.128(0.000)
$\Delta_f G_{NADP}$	-2030899.087(0.000)	-2030899.560(0.000)	-2030900.850(0.000)	-2030899.704(0.000)
$\Delta_f G_{GL6P}$	-1376900.009(0.000)	-1376901.145(0.000)	-1376900.443(0.000)	-1376900.755(0.000)
$\Delta_f G_{NADPH}$	-1964198.964(0.000)	-1964198.421(0.000)	-1964198.265(0.000)	-1964199.013(0.000)
$\Delta_f G_{PGn}$	-1553399.717(0.000)	-1553399.686(0.000)	-1553399.111(0.000)	-1553399.776(0.000)
$\Delta_f G_{Ru5P}$	-1223400.056(0.000)	-1223399.774(0.000)	-1223400.149(0.000)	-1223399.864(0.000)
$\Delta_f G_{CO2}$	-403099.977(0.000)	-403100.201(0.000)	-403100.248(0.000)	-403100.161(0.000)
$\Delta_f G_{R5P}$	-1225401.343(0.000)	-1225400.034(0.000)	-1225400.981(0.000)	-1225400.031(0.000)
$\Delta_f G_{X5P}$	-1226798.954(0.000)	-1226799.735(0.000)	-1226798.810(0.000)	-1226799.272(0.000)
$\Delta_f G_{E4P}$	-1155799.976(0.000)	-1155800.272(0.000)	-1155801.314(0.000)	-1155800.240(0.000)
$\Delta_f G_{S7P}$	-1364300.161(0.000)	-1364299.544(0.000)	-1364300.041(0.000)	-1364299.739(0.000)
$\Delta_f G_{KDPG}$	-1444899.661(0.000)	-1444900.590(0.000)	-1444901.283(0.000)	-1444899.969(0.000)
$\Delta_f G_{CoA}$	-1719298.837(0.000)	-1719299.584(0.000)	-1719299.772(0.000)	-1719299.176(0.000)
$\Delta_f G_{AcCoA}$	-1772299.229(0.000)	-1772299.135(0.000)	-1772300.469(0.000)	-1772299.127(0.000)
$\Delta_f G_{Phosph}$	-1070600.257(0.000)	-1070599.216(0.000)	-1070600.219(0.000)	-1070599.137(0.000)
$\Delta_f G_{AcCoA}$	-1772299.599(0.000)	-1772298.796(0.000)	-1772299.824(0.000)	-1772299.352(0.000)
$\Delta_f G_{CoA}$	-1719298.640(0.000)	-1719301.057(0.000)	-1719300.974(0.000)	-1719301.355(0.000)
$\Delta_f G_{Ace}$	-255000.089(0.000)	-255000.130(0.000)	-254999.965(0.000)	-255000.200(0.000)
$\Delta_f G_{OAA}$	-726299.797(0.000)	-726300.391(0.000)	-726299.658(0.000)	-726300.337(0.000)
$\Delta_f G_{Cit}$	-969599.697(0.000)	-969600.130(0.000)	-969600.498(0.000)	-969599.912(0.000)
$\Delta_f G_{Icit}$	-961999.542(0.000)	-961999.812(0.000)	-962000.954(0.000)	-961999.867(0.000)
$\Delta_f G_{aKg}$	-636800.236(0.000)	-636799.874(0.000)	-636799.739(0.000)	-636800.053(0.000)
$\Delta_f G_{SucCoA}$	-2048298.882(0.000)	-2048299.844(0.000)	-2048301.149(0.000)	-2048299.435(0.000)
$\Delta_f G_{Suc}$	-524300.075(0.000)	-524300.465(0.000)	-524299.906(0.000)	-524300.412(0.000)
$\Delta_f G_{Ubi}$	774899.412(0.000)	774899.979(0.000)	774900.588(0.000)	774899.965(0.000)
$\Delta_f G_{Fum}$	-531400.135(0.000)	-531400.233(0.000)	-531399.605(0.000)	-531400.245(0.000)
$\Delta_f G_{UbiH2}$	760400.042(0.000)	760398.968(0.000)	760399.315(0.000)	760399.056(0.000)
$\Delta_f G_{Mal}$	-686401.360(0.000)	-686400.222(0.000)	-686400.764(0.000)	-686400.208(0.000)

C.1 Analytical derivation of the relaxation time

First, let us derive the general solution of the differential equation governing a set of concentrations $x_m(t)$ $(m = 1,...,N)$:

$$\frac{d\,x_m}{d\,t} = Ax_{m-1} - (A+B)x_m + Bx_{m+1} \tag{C.1}$$

Separating the dependence on time and position, we set $x_m(t) = T(t)X_m$. Then

$$\frac{T'}{T} = \frac{AX_{m-1} - (A+B)X_m + BX_{m+1}}{X_m} \tag{C.2}$$

The left hand side is independent of m while the right hand side is independent of t, so both must be equal to the same constant which we denote by $-\lambda$.

The dependence on time is immediate: $T(t) = T(0)\exp(-\lambda t)$.

The equations eq.C.2 for the X_m correspond to finding an eigenvector of a matrix H of eigenvalue $-\lambda$. The diagonal elements of H are $-(A+B)$ while the off diagonal parts are restricted to nearest neighbors and are again independent of m. Given this translational invariance, the eigenvectors can be taken to be of the form $X_m = X_0 e^{(\gamma \pm j\omega)m}$. Plugging this in, one has

$$-\lambda = Ae^{-(\gamma \pm j\omega)} - (A+B) + Be^{-(\gamma \pm j\omega)}$$

For the studied system, both A and B are real.

Let us start by searching for the real eigen values. We can solve for the real and imaginary parts:

Imaginary part	Real part
$Ae^{-\gamma}\sin(\mp\omega) + Be^{\gamma}\sin(\pm\omega) = 0$	$\lambda = Ae^{-\gamma}\cos(\omega) - (A+B) + Be^{\gamma}\cos(\omega)$
$\gamma = \frac{1}{2}\log\left(\frac{A}{B}\right)$	$\lambda = 2\sqrt{AB}\cos(\omega) - (A+B)$

A general solution of eq.C.2 is a combination of the two complex conjugates X_m^+ and X_m^-:

$$X_m = C^- X_m^- + C^+ X_m^+ = C^- e^{(\gamma - j\omega)m} + C^+ e^{(\gamma + j\omega)m}$$

Whith boundary conditions $X_0 = X_{N+1} = 0$:

$X_0 = 0$

$$\begin{aligned} 0 &= C^- + C^+ \\ C^- &= -C^+ \\ X_m &= Ce^{\gamma m}\sin(\omega m) \end{aligned}$$

$X_{N+1} = 0$

$$\begin{aligned} \sin(\omega(N+1)) &= 0 \\ \omega_\kappa &= \frac{\kappa\pi}{N+1} \quad \kappa \in \{0, 1, ..., N-1\}. \end{aligned}$$

The search for real eigen values provides us with N real eigen values which is the rank of the matrix H, thus all the eigen values are real.

$$\begin{aligned} x_m^\kappa(t) &= C^\kappa e^{\frac{1}{2}\log(A/B)m}\sin\left(\frac{m\kappa\pi}{N+1}\right)e^{\lambda_\kappa t} \\ &\text{with} \\ \lambda_\kappa &= 2\sqrt{AB}\cos\left(\frac{\kappa\pi}{N+1}\right) - (A+B), \quad \kappa \in \boldsymbol{Z}^* \end{aligned} \tag{C.3}$$

A relaxation time is associated to each λ_κ:

$$\tau_\kappa = \frac{1}{A + B - 2\sqrt{AB}\cos\left(\frac{\kappa\pi}{N+1}\right)} \tag{C.4}$$

In practice, one refers to *the* relaxation time as the longest such times, *i.e.*, $\tau = \frac{1}{A+B-2\sqrt{AB}\cos\left(\frac{\pi}{N+1}\right)}$.

Case of a symmetric system: For a system that is left-right symmetric ($A = B$), $\tau \approx \frac{(N+1)^2}{A\pi^2}$ for large N which is characteristic of diffusing systems. (Clearly, if $A = B$, the equations describe a diffusing particle that has no bias towards left or right.)

Case of an asymmetric system: If the system is not left-right symmetric ($A \neq B$), at large N the relaxation time become independent of N (and in fact at fixed κ all τ_κ become independent of N:

$$\tau_\kappa \approx -\frac{1}{2\sqrt{AB} - (A+B)}$$

.

A crossover size for N between the two regimes $\tau \propto N^2$ and $\tau = cst$: If the difference between A and B is sufficiently small, the relaxation time will behave as the case $A = B$, that is τ will scale as N^2 until N is large enough for the regime of constant τ to set in. To understand the scale in N where this crossover behavior arises, let us note $A = B + \epsilon$. Then for large N

$$\begin{aligned}
1/\tau &\approx 2B + \epsilon B - 2B\sqrt{1+\epsilon}(1 - \frac{\pi^2}{2(N+1)^2}) \\
1/\tau &\approx 2B + \epsilon B - 2B(1 + \frac{\epsilon}{2} - \frac{\epsilon^2}{8})(1 - \frac{\pi^2}{2(N+1)^2}) \\
1/\tau &\approx 2B + \epsilon B - 2B - B\epsilon + \frac{B\epsilon^2}{4} + \frac{B\pi^2}{(N+1)^2} + \frac{B\epsilon\pi^2}{2(N+1)^2} \\
1/\tau &\approx \frac{B\epsilon^2}{4} + \frac{B\pi^2}{(N+1)^2}
\end{aligned}$$

This last equation shows that one has a crossover between the two regimes when

$$\begin{aligned}
\frac{B\epsilon^2}{4} &\approx \frac{B\pi^2}{(N+1)^2} \\
\frac{\epsilon}{2} &\approx \frac{\pi}{(N+1)}
\end{aligned}$$

The transition between a diffusing regime and a regime independent on the network size N thus occurs for a characteristic size

$$N^{cross} = 1 + \frac{2B\pi}{A - B} \tag{C.5}$$

D

Description of the algorithms used in the thesis

D.1 Computing the different characteristic times

Our characteristic times are defined using infinitesimal perturbations of the steady state. Such a choice has the advantage of facilitating the treatment of the time dependence of the perturbation since one can exploit the linearity of the dynamics. Denoting by $\vec{C}$ the infinitesimal vector giving the concentration deviation from the steady state and by $\mathbf{J}$ the jacobian matrix associated with the linearized dynamics, one has:

$$\begin{aligned} \frac{d\vec{C}}{dt} &= \mathbf{J}\vec{C} \\ \vec{C}(t) &= e^{\mathbf{J}t}\vec{C}(t=0) \\ |\vec{C}(t)| &= |e^{\mathbf{J}t}\vec{C}(t=0)| \end{aligned}$$

In Annex C.1 we show that the eigenvalues of $\mathbf{J}$ are all real and negative in the case of a homogeneous linear chain as might be expected in a dissipative system.

Our matrix formulation allows us to conveniently define the "lifetime" associated with a perturbation $\vec{C}(t)$ introduced at time $t = 0$ via

$$\begin{aligned} T &= \frac{1}{|\vec{C}(t=0)|}\int_0^\infty |\vec{C}(t)|dt \\ &= \frac{1}{|\vec{C}(t=0)|}\int_0^\infty |e^{\mathbf{J}t}\vec{C}(t=0)|dt \end{aligned}$$

Because of the linearity, the characteristic time T does not depend on the *amplitude* of the initial perturbation, $|C(\vec{t=0})|$. In addition, we shall impose the initial perturbation to be localized to just one site of the metabolic network. In consequence, and without any loss of generality, we can take the initial perturbation to be positive, and then it is easy to see that all of the components of $\vec{C(t)}$ remain positive at all times. using this property, we can interchange the integration and the taking of the norm if we work with the L_1 norm:

$$\begin{aligned} T &= \frac{1}{|\vec{C}(t=0)|}|\int_0^\infty e^{\mathbf{J}t}\vec{C}(t=0)dt| \\ &= \frac{1}{|\vec{C}(t=0)|}|[\mathbf{J}^{-1}e^{\mathbf{J}t}\vec{C}(t=0)]_0^\infty| \\ &= \frac{1}{|\vec{C}(t=0)|}|\mathbf{J}^{-1}\vec{C}(t=0)| \end{aligned}$$

The lifetime T as specified is dependent on the position where the perturbation was introduced at $t = 0$. To overcome this drawback, we define the characteristic time of the system as the maximal lifetime when considering all possible positions of the initial perturbation.

Our global algorithm to compute our characteristic time is then as follows:

Data: vector_param: best_vector,conc_ss: concentration at steady state, F: function returning the derivative the system
Result: lifetime of the system
J = linearize F(...,param) near conc_ss;
max_time = 0;
for *i=1 to i=conc_ss* **do**
  $\vec{C}^0$ = {0,...,0};
  $\vec{C}^0$[i] = 1;
  time = $\frac{1}{|\vec{C}(t=0)|}|\mathbf{J}^{-1}\vec{C}^0|$;
  if *time>max_time* **then**
    max_time = time;
  end
end
return *max_time*;

Algorithm 1: Algorithm to compute the lifetime

D.2 Finding a model with a satisfactory steady state

Before optimizing the parameters, one first needs to find an initial model for which there is a "satisfactory" steady state, namely one where all the fluxes are strictly positive with respect to the reference direction of the fluxes. Said differently, we want the ratio Q/k_{eq} must be smaller than 1 for every reaction. If that is not imposed, it turns out to be far more difficult to adjust the parameters to have the model's behavior approach the desired target. I provide here the logic of my algorithm for obtaining

such an initial model:

Data: All the paremeters are centred on their reference distribution
Result: A set of parameters leading to positive fluxes
Search initial steady state;
while *Steady state is not satisfactory* **do**
  Compute quotients of reaction, Q from conc ;
  Score_old = $\sum_i \frac{Q^i}{k_{eq}^i}$;
  save_conc = conc;
  while *Any* $Q/k_{eq} > 1$ **do**
    i = Position(max(Q/k_{eq}));
    for *j=0 to length(conc)* **do**
      if *conc[j] reacts through reac[i]* **then**
        conc[j] += $(1 + \delta \times \frac{d\,reac[i]}{d\,conc[j]})$;
      end
    end
    Compute quotients of reaction, Q from conc;
    Score_new = $\sum_i \frac{Q^i}{k_{eq}^i}$;
    if *Score_new>Score_old* **then**
      conc = save_conc;
    end
  end
  for *i=1 to i=length(reac)* **do**
    $V_m^i \leftarrow V_m^i \times \frac{v^i}{v^{i,ref}}$;
  end
  Perturb the new steady state and search the new steady state to test the stability;
end

Algorithm 2: Search for satisfactory initial model

D.3 Optimization algorithm

It is necessary to optimize the parameters so that the model satisfies at best the different constraints imposed. The simplest optimization algorithms rely on "iterated improvement". In such algorithms, the vector of parameters is successively modified stochastically, for instance by changing one or a small number of parameters by some small random percentage. For each such modification, the score of the modified model is evaluated. If the score is higher than the previous one so that one has found an improvement, one accepts the modifications; otherwise one discards them. After repeating these steps some number of times, thereby iteratively improving the model, at some point the score can no longer be improved and one has reached a local or global maximum of the score.

When applying a modification to the parameters, it may happen that some of the resulting steady state fluxes go in the wrong direction with respect to the reference fluxes. In this case, the score is not even computed and I discard the modifications.

The iterative improvement approach is simple but in the context of my work it turned out to be quite inefficient. Perhaps that inefficiency is associated with the large number of parameters in my system and to the sensitivity of the score to independent changes of the parameters. To overcome this obstacle, I implemented a genetic algorithm to better optimize the score. This algorithm is based on following a population of models across successive generations. Each model in the population has its parameters modified as for the iterated improvement approach, but instead of forcing an improvement of the score

I simply keep the best models at each generation.

Data: init_param: Initial parameter vector,function_score: function that evaluate the score associated to a parameter vector, δ: vector of perturbation amplitude
Result: Parameter vector leading to the optimal steady state for a set of constraints
Search initial steady state;
score = function_score(param);
no_improvement = 0;
N = Number of parameter vectors;
list_param = create an empty list of parameter vectors;
best_vector = init_param;
while *length(list_param)<N* **do**
  temp_vector = best_vector;
  for *i=1 to i=5* **do**
    param_to_perturb = select uniformly an parameter index;
    temp_vector[param_to_perturb] *= δ[param_to_perturb]*random_gauss($\mu = 0, \sigma = 1$);
    if *temp_vector produces a satisfactory model* **then**
      Add temp_vector to list_param;
    end
  end
end
while *no_improvement < 50* **do**
  while *length(list_param)<2N* **do**
    parent1,parent2 = Select uniformly 2 parents;
    temp_vector = list_param[parent1];
    for *i=1 to i=nb_param-1* **do**
      if *random_unif*($\{0,1\}$) *is 1* **then**
        temp_vector[i] = list_param[parent2][i]
      end
    end
    if *temp_vector is satisfactory* **then**
      Add temp_vector to list_param;
    end
  end
  Evaluate the score for every parameter vector in list_param;
  list_param = Select N-1 vectors with the best scores;
  if *Any vector in list_param is better than best_vector* **then**
    best_vector = new best vector in the list_param;
  else
    no_improvement ← no_improvement+1;
  end
  Add best_vector to list_param;
end
return *best_vector*;

Algorithm 3: Genetic algorithm for parameter optimization

Not only was this genetic algorithm more efficient (leading to faster improvement in the score), but it also did not get stuck in local maxima. Nevertheless, it turns out that the optimization led to models with long characteristic times. This is undesirable for two reasons: first, the biological situation probably does not have large relaxation times, and second, having such large relaxation times led to numerical difficulties in determining the steady states. To avoid this problem, I set a maximum value on the relaxation time. When it went beyond 1000 sec, I discarded the model with the modified parameters, regardless of the associated score.

D.4 Estimating confidence intervals for the parameters

The confidence intervals for the parameters are computed from an ensemble of models generated through Markov Chain Monte Carlo (MCMC). The sample of models produced by MCMC is obtained from a kind of random walk in the space of all models, based on the following pseudo-code:

Data: vector_param: best_vector,function_score: function that evaluate the score associeted to a parameter vector, δ: vector of perturbation amplitude
Result: Parameter vector leading to the optimal steady state for a set of constraints
Search initial steady state;
score = function_score(param);
list_param = create an empty list of parameter vectors;
while *length(list_param)*<*N* **do**
    temp_vector = vector_param;
    for *i=1 to i=5* **do**
        param_to_perturb = select uniformly an parameter index;
        temp_vector[param_to_perturb] *= δ[param_to_perturb]*random_unif([−1…1]);
    end
    if *temp_vector produces an unsatisfactory model* **then**
        continue;
    end
    score_new = function_score(temp_vector);
    if *random_unif(*[0…1])> $e^{score\_new - score}$ **then**
        Add temp_vector to list_param;
        vector_param = temp_vector;
    end
end
return *list_param*;

Algorithm 4: MCMC algorithm for for sampling parameters

D.5 Programs used in the thesis

The programs succinctly described in these annexes and used extensively in this thesis are available at the following location:

https://github.com/adrienhenry/source_code_thesis.git/